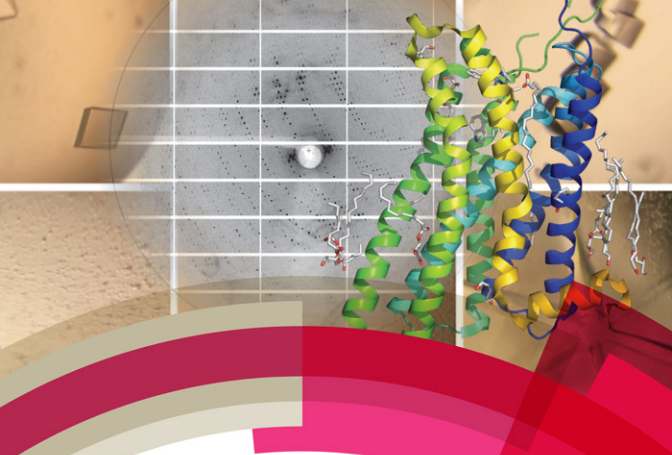


Chemical Biology

Protein Crystallography

Challenges and Practical Solutions

Konstantinos Beis and Gwyndaf Evans

ROYAL SOCIETY OF CHEMISTRY

Protein Crystallography
Challenges and Practical Solutions

Chemical Biology

Editor-in-chief:
Tom Brown, *University of Oxford, UK*

Series editors:
Kira J. Weissman, *Lorraine University, France*
Sabine Flitsch, *University of Manchester, UK*
Nick J. Westwood, *University of St Andrews, UK*

Titles in the series:
1: High Throughput Screening Methods: Evolution and Refinement
2: Chemical Biology of Glycoproteins
3: Computational Tools for Chemical Biology
4: Mass Spectrometry in Chemical Biology: Evolving Applications
5: Mechanisms of Primary Energy Transduction in Biology
6: Cyclic Peptides: From Bioorganic Synthesis to Applications
7: DNA-targeting Molecules as Therapeutic Agents
8: Protein Crystallography: Challenges and Practical Solutions

How to obtain future titles on publication:
A standing order plan is available for this series. A standing order will bring delivery of each new volume immediately on publication.

For further information please contact:
Book Sales Department, Royal Society of Chemistry, Thomas Graham House, Science Park, Milton Road, Cambridge, CB4 0WF, UK
Telephone: +44 (0)1223 420066, Fax: +44 (0)1223 420247,
Email: booksales@rsc.org
Visit our website at www.rsc.org/books

Protein Crystallography
Challenges and Practical Solutions

Edited by

Konstantinos Beis
Imperial College London, UK
Email: kbeis@imperial.ac.uk

and

Gwyndaf Evans
Diamond Light Source, UK
Email: gwyndaf.evans@diamond.ac.uk

Chemical Biology No. 8

Print ISBN: 978-1-78262-728-9
PDF ISBN: 978-1-78801-050-4
EPUB ISBN: 978-1-78801-477-9
Print ISSN: 2055-1975
Electronic ISSN: 2055-1983

A catalogue record for this book is available from the British Library

© The Royal Society of Chemistry 2018

All rights reserved

Apart from fair dealing for the purposes of research for non-commercial purposes or for private study, criticism or review, as permitted under the Copyright, Designs and Patents Act 1988 and the Copyright and Related Rights Regulations 2003, this publication may not be reproduced, stored or transmitted, in any form or by any means, without the prior permission in writing of the Royal Society of Chemistry or the copyright owner, or in the case of reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency in the UK, or in accordance with the terms of the licences issued by the appropriate Reproduction Rights Organization outside the UK. Enquiries concerning reproduction outside the terms stated here should be sent to the Royal Society of Chemistry at the address printed on this page.

Whilst this material has been produced with all due care, The Royal Society of Chemistry cannot be held responsible or liable for its accuracy and completeness, nor for any consequences arising from any errors or the use of the information contained in this publication. The publication of advertisements does not constitute any endorsement by The Royal Society of Chemistry or Authors of any products advertised. The views and opinions advanced by contributors do not necessarily reflect those of The Royal Society of Chemistry which shall not be liable for any resulting loss or damage arising as a result of reliance upon this material.

The Royal Society of Chemistry is a charity, registered in England and Wales, Number 207890, and a company incorporated in England by Royal Charter (Registered No. RC000524), registered office: Burlington House, Piccadilly, London W1J 0BA, UK, Telephone: +44 (0) 207 4378 6556.

For further information see our web site at www.rsc.org

Printed in the United Kingdom by CPI Group (UK) Ltd, Croydon, CR0 4YY, UK

FOREWORD

Protein Crystallography: Faster, Smaller, Stronger

DAVID I. STUART* AND JAMES H. NAISMITH

*Email: dave@strubi.ox.ac.uk

The symmetry and beauty of crystals, found commonly as minerals in rocks and of course in the production of salt, has long been a source of fascination. The underlying determinant of crystal morphology emerged when the work of Ernest von Laue revealed that crystals diffracted X-rays and Lawrence Bragg working alongside his father William derived a formula which explained the phenomenon. Since X-rays were being scattered by electrons, it was realised that these diffracted beams held clues to the underlying arrangements of the atoms and thereby the structure of the material. Lawrence Bragg in his Nobel Lecture noted ''The examination of crystal structure, with the aid of X-rays has given us for the first time an insight into the actual arrangement of the atoms in solid bodies.'' Presciently, he noted ''There seems to be hardly any type of matter in the condition of a true solid which we cannot attempt to analyse by means of X-rays.'' In these early days, the structures 'determined' by crystallography were simple salts. Since it was already known that many organic and inorganic compounds could be crystallised, the need to develop robust methods to recover the relative phases of the diffracted beams was plain, if more complex structures were to be solved. Progress continued, with notable early landmarks in 1929; Kathleen Lonsdale's proof that benzene rings were flat and John (JD) Bernal's correct structure of cholesterol. Almost simultaneously, James B. Sumner had established beyond doubt that enzymes were a single defined chemical species that could be purified

Chemical Biology No. 8
Protein Crystallography: Challenges and Practical Solutions
Edited by Konstantinos Beis and Gwyndaf Evans
© The Royal Society of Chemistry 2018
Published by the Royal Society of Chemistry, www.rsc.org

the same as any other chemical substance. Sumner grew crystals of urease and concanavalin A isolated from jack bean; his work inspired others to use crystallisation as an increasingly routine means of purifying proteins (and establishing their purity). Significant numbers of proteins were being crystallised by biochemists.

The advances in these fields of crystallography and protein biochemistry continued essentially in parallel. Up to the development of direct methods begun in 1953 by Herbert Hauptman and Jerome Karl, the solution of so-called small molecule crystal structures involved time consuming Patterson methods, placement of known fragments, isomorphous series, heavy atoms and intuition. As direct methods expanded in power and utility, the solving of the structure of small molecules by crystallography became increasingly routine. The process was dramatically accelerated by the increasing availability and power of computers. The power of supercomputers was seen in the early 1950's when Dorothy Crowfoot Hodgkin and Jenny Glusker, working on the structure of vitamin B12, called in Ken Trueblood at UCLA to accelerate the determination of the structure by 'post and cable' support from his local machine (which was of course millions of times slower than a smartphone).

From its inception, protein crystallography has involved physicists, chemists and biologists working at the interface of their disciplines. Working with J. D. Bernal, Hodgkin in 1934 showed that protein crystals, like those of common salt, diffracted X-rays according to Bragg's law. A key experimental constraint was that the protein crystals had to be kept hydrated. Max Perutz, a PhD student encouraged by both J. D. Bernal and Lawrence Bragg, began to study crystals of haemoglobin by X-ray diffraction. Interrupted by the Second World War, Perutz resumed his work and showed that by replacing a water molecule with a much heavier atom such as mercury, it was possible to experimentally determine the phases and thus the structure of haemoglobin. His colleague John Kendrew used the same approach to determine the structure of myoglobin. In 1959, these structures were available around 5.5 and 2 Å respectively and provided unprecedented insights into the molecular mechanism of oxygen binding and transport.

Others followed suit, Hodgkin, already a Nobel Prize winner for her work on penicillin and vitamin B12, determined the structure of insulin in 1969, David Philips lysozyme in 1965 and David Blow chymotrypsin in 1967. In each case, these structures revolutionised our understanding of the molecular basis of biology. As Perutz said, ''I think we pushed knowledge of living matter down to the atomic level and this enormously increased our depths of understanding of the nature of life.'' This power was widely accepted by 1970. A method for structure solution using related structures emerged in the early 1970's. This, the molecular replacement method, is now used for most structural analyses of new protein structures and was developed following first suggestions by Walter Hoppe, and then Michael Rossmann and David Blow. Viruses, DNA and RNA oligomers, and

membrane proteins all were shown to be crystallisable and thus be amenable to study. The simple act of crystallising viruses was a major landmark in virology. In 1935 Wendell Stanley crystallized tobacco mosaic virus, which led to a share of a Nobel prize, the first awarded to virology. This of course was many years before structure determination was feasible, but it was clear that Bragg's prediction of the ubiquity of crystal structure analysis was true. During the 1970's the technique became widely established, although new structures came slowly.

The utility of protein crystallography in the 1970's and early 1980's was limited by significant technical challenges. Structure solutions were hard won following years of trial and error experiments; data collection moved to photographic film, which was time consuming; X-ray sources were weak (access to powerful synchrotrons was not yet routine); only the most abundant proteins were plausible candidates for isolation; many proteins only gave small crystals and when these crystals could be examined at synchrotrons they often died during data collection. In a period of about five years from the mid 1980's each of these problems was overcome by what individually could be characterised as incremental advances but when combined were transformative. Perhaps the greatest single impact was the application of recombinant DNA technology. The principles were discovered in the 1970's and the development of methods over subsequent years eventually opened up even the most recalcitrant human proteins to crystallographic analysis. This allowed structural biologists to focus on the biological significance of a target, rather than its natural abundance. Steady advances in molecular biology now make overexpression and simple purification routine, with the structural genomics/proteomics post-millennium efforts putting structural biologists centre stage in the development of protein production methods. Freezing of crystals before data collection greatly reduced their decay during exposure to X-rays improving data quality, new beamlines at synchrotrons came on stream giving much more access, machine readable image plates replaced film greatly speeding up and improving data collection and the power and availability of small computers meant that even small groups could solve structures quickly in their own laboratory. Increased intensity and better focus of synchrotron beams made it possible to study much smaller crystals. The replacement of methionine with selenomethionine, pioneered by Wayne Hendrickson, became a generic vehicle, making phase determination using anomalous scattering signals almost routine, accelerated by access to tuneable high energy synchrotron beams (in the UK this was helped by the MRC's purchase, with EMBL, of BM14 at the ESRF). In the 1990's the number of structures being determined exploded. By this point all major pharmaceutical companies were using the technique in drug discovery and most biological problems were being advanced by structural study at some level. In the years following 2000, industry and structural genomics/proteomics consortia drove a further round of advances, notably standardisation which facilitated automation, remote access and near real

time data analysis. Then in 2011 a new chapter opened with the first publications of results from a hard X-ray free electron laser (XFEL), the ultimate impact of XFEL on protein crystallography is still an open question, and is discussed in the current volume.

The power and scope of the technique continues to grow and although cryo electron microscopy has dethroned crystallography for larger assemblies, it remains crucial for many different approaches. This timely book highlights some of the key challenges protein crystallography faces and opportunities for its expansion. True to the technique's origins, the authors span a range of scientific disciplines. In Chapter 1 **Martiel *et al.*** discuss *in situ* crystallography, a term that signals a renaissance in room temperature crystallography, focussed on avoiding the problematic stages of crystal handling and mounting, and is now offered in various forms at many synchrotrons and XFELs. In the following chapter **Abola *et al.*** develop this theme to focus on one of the major success stories of XFELs, the determination of the structure of G-protein coupled receptors. Free electron lasers with their brilliance and pulse time structure are beginning to make an impact driving new techniques, as described here. This chapter reminds us of the ultimate goal of the technique, to make a difference to human health. Molecules that target GPCR's account for 40% of medicines in use but until the structural insights in 2007, the rational design of such molecules was limited to models based on rhodopsin. The past ten years have seen multiple new drug companies emerge using the structure of GPCR's to develop new medicines. **Foos *et al.***, return to the problem addressed by Martiel of how to further optimise the use of synchrotron radiation, this time looking at ways of optimising the collection of fragmentary cryo data sets; methods such as this should ultimately allow high quality data sets to be collected rapidly as slivers and assembled together automatically from very large numbers of microcrystals. **Garman and Weik** bring the reader up to date on the decay of protein crystals upon X-ray exposure, a problem which although diminished by cryo-crystallography is far from eliminated. Although our knowledge is still incomplete, the authors demonstrate how our present understanding of the problem has led to approaches to reduce its baleful consequences and just occasionally how it can work to our advantage. The central importance of extracting the most accurate and meaningful data from experiments is stressed by **Vollmar and Evans** who set out improvements in data processing (speed, depth of information and accuracy) and provide a useful checklist for collecting good quality data. The volume continues with two chapters illustrating testing examples; **Symersky *et al.*** describe improved methodologies for working with low-resolution, anisotropic data and crystal twinning while **Bunker's** work shows that even low-resolution data can transform our understanding of biology. Both of these contributions are reminders that although technological advances make crystallography sometimes seem almost facile, accessing interesting biology can require painstaking work, and such work highlights weaknesses in our current

arsenal of techniques – it should by now be routine to be able to refine structures against a target function derived from multiple non-isomorphous datasets. We hope that such challenging projects will drive the next round of methods development. Finally, **Zatsepin** provides an in-depth review on recent advances in XFELs from sample delivery to data processing. She highlights the advantages of using XFELs to study challenging membrane proteins as well as to study enzymatic reactions that would not be possible in conventional synchrotrons.

Finally, we should remind ourselves that the early pioneers of the field were sharply focussed on problems of medical relevance; Hodgkin solved penicillin, vitamin B12 and insulin, Perutz saw the structure of haemoglobin as a way to explain its function and mis-function and of course lysozyme is a natural anti-bacterial defence mechanism. Now that macromolecular crystallography is a mature subject, capable of delivering a structure from a well-behaved crystal in minutes, the challenge is to make the most of the method. Hopefully this book might contribute by helping some of the very large number of practitioners to understand the opportunities and problems, and perhaps avoid some pitfalls.

Contents

Chapter 1 **Practical Approaches for *In Situ* X-ray Crystallography: from High-throughput Screening to Serial Data Collection** **1**
Isabelle Martiel, Vincent Olieric, Martin Caffrey and Meitian Wang

1.1 Introduction 1
1.1.1 What Exactly Is *In Situ*? 1
1.1.2 Goals of *In Situ* Experiments 3
1.1.3 Challenges of *In Situ* Methods 4
1.1.4 Enabling Technologies 5
1.2 *In Situ* Screening at the Synchrotron: Standard SBS Plates 7
1.2.1 Development History 7
1.2.2 Plate Handling Hardware 9
1.2.3 Plate Optimization for *In Situ* 9
1.2.4 Automation and Pipeline Integration 10
1.3 Further Developments: Scale Reduction and Microfluidics 11
1.3.1 Small Formats 11
1.3.2 Microfluidic Methods for *In Situ* 13
1.4 The Emergence of Serial *In Situ* Data Collection 15
1.4.1 Thin-film Sandwiches 15
1.4.2 Liquid Manipulation Methods 19
1.5 Conclusion and Outlook 20
Acknowledgements 20
References 21

Chemical Biology No. 8
Protein Crystallography: Challenges and Practical Solutions
Edited by Konstantinos Beis and Gwyndaf Evans
© The Royal Society of Chemistry 2018
Published by the Royal Society of Chemistry, www.rsc.org

Chapter 2 Delivery of GPCR Crystals for Serial Femtosecond Crystallography 28
E. E. Abola, U. Weierstall, W. Liu and V. Cherezov

2.1 Introduction 28
2.2 Process Overview 30
2.3 Achieving Major Milestones 31
2.3.1 Large-scale Production of Stable Receptor Constructs 31
2.3.2 Crystallization of Receptor Constructs 38
2.3.3 SFX Data Collection 39
2.4 Summary of Successful GPCR Structural Studies at XFELs 42
Acknowledgements 46
References 46

Chapter 3 The Mesh&Collect Pipeline for the Collection of Multi-crystal Data Sets in Macromolecular Crystallography 54
Nicolas Foos, Gleb Bourenkov, Gordon Leonard, Igor Melnikov, Christoph Mueller-Dieckmann, Max Nanao, Alexander Popov, Gianluca Santoni and Ulrich Zander

3.1 Introduction 54
3.2 Dozor 59
3.3 Hierarchical Cluster Analysis (HCA) 59
3.4 Mesh&Collect in Practice 60
3.4.1 Solving the Crystal Structures of Membrane Proteins with Very Small Crystals 61
3.4.2 Multi-crystal Data Collection for Ligand Binding Studies 63
3.4.3 *De novo* Structure Solution using Mesh&Collect 64
3.4.4 Mesh&Collect for the I-SAD/I-SIRAS Solutions of the Crystal Structure of the KR2 Light-Driven Sodium Pump 68
3.4.5 Mesh&Collect at Room Temperature 69
3.5 The Pitfalls of HCA 69
3.6 ccCluster 75
3.7 Merging of Partial Data Sets Using Genetic Algorithms 75
3.7.1 Grouping Partial Data Sets into Chromosomes 77
3.7.2 Fitness Evaluation 77

3.7.3 GA Optimisation 79
3.7.4 Case Study: LUX 79
3.8 MeshBest 80
3.8.1 NarQ Crystals Analysed by a Mesh Scan 82
3.8.2 A 'Mishmash' of Thaumatin Crystals 82
3.9 Conclusions 83
References 84

Chapter 4 Radiation Damage in Macromolecular Crystallography 88
Elspeth F. Garman and Martin Weik

4.1 Introduction 88
4.2 How Do X-ray Photons Interact with Matter? 90
4.3 Global and Specific Radiation Damage Effects at 100 K and Below 92
4.4 Estimating the Absorbed Dose and Dose Limits 96
4.5 X-ray Induced Changes in Chromophore-containing Proteins at 100 K 102
4.6 Global and Specific Radiation Damage Above 100 K and at Room Temperature 102
4.7 Recruitment of Radiation-induced Changes to Study Macromolecular Function 103
4.8 Radiation-damage Induced Phasing 104
4.9 Does Radiation Damage Depend on Dose Rate and/or on the Incident Beam Energy at 100 K? 105
4.10 How Can Radiation Damage Be Minimised? 106
4.11 Radiation Damage in Serial Femtosecond Crystallography at XFELs 108
Acknowledgements 109
References 109

Chapter 5 Data Quality Analysis 117
M. Vollmar and Phil Evans

5.1 Introduction 117
5.2 Accuracy *versus* Precision and Merged *versus* Unmerged Data 118
5.3 Sources of Error 119
5.3.1 Random Error 120
5.3.2 Systematic Error 121
5.3.3 Outliers 123
5.3.4 Radiation Damage 124

5.4 Estimating Errors 125
5.4.1 Estimation of $\sigma(I_{hkl})$ 125
5.4.2 *ISa*, an Indicator for Systematic Error 126
5.4.3 R_{merge}, R_{sym} and R_{meas} – Indicators for Unmerged Data 126
5.4.4 $R_{mrgd\text{-}I}$, $R_{p.i.m.}$, R_{anom} and $CC_{1/2}$ – Indicators for Merged Data 127
5.4.5 R_d, R_{cum} and *B Factor* – Indicators for Radiation Damage 129
5.4.6 Data Completeness 130
5.5 Use of Metrics 130
5.5.1 BLEND and Merging Multiple Crystals and/or Data Sets 130
5.5.2 Identifying Rogue Data Sets and Linking Data and Model Quality 131
5.5.3 Determination of a High-resolution Cut-off 132
5.5.4 Things to Consider When Collecting and Analysing Data 134
5.6 Concluding Remarks 137
References 138

Chapter 6 Structure Determination at Low-resolution, Anisotropic Data and Crystal Twinning 140
Jindrich Symersky, Yi Guo, Jimin Wang and Min Lu

6.1 Introduction 140
6.2 Experimental 141
6.2.1 Protein Expression and Purification 141
6.2.2 Protein Crystallization and Derivatization 142
6.2.3 X-ray Diffraction Data Collection and Analysis 142
6.3 Results and Discussion 142
6.3.1 Data Collection and Radiation Damage 142
6.3.2 Data Reduction and Anisotropy 143
6.3.3 Twinning Detection and Analysis 144
6.3.4 Molecular Replacement and Structure Solution 147
6.3.5 Difference Fourier Analysis 149
6.3.6 Attempts in Experimental Phasing 151
6.3.7 Comparison of the NorM-NG Structures 152
6.3.8 Ligand-binding Site 153

6.4 Conclusion 154
Acknowledgements 155
References 155

Chapter 7 Structure Determination and Refinement of Large Macromolecular Assemblies at Low Resolution 157
Richard D. Bunker

7.1 Introduction 157
7.2 Crystallization, Data Collection and Processing 159
7.3 Crystal Characterisation 161
7.4 CSN4 165
7.5 Heavy-atom-soaked Derivative Crystals 165
7.6 Initial Phasing 166
7.7 Subunit Identification and Selenomethionine Phasing 169
7.8 Initial Model Building 170
7.9 Model Completion 171
7.10 Analysis of CSN Conformational Dynamics Aided by a *P*1 Crystal Form 174
7.11 Conclusions 177
Acknowledgements 178
References 178

Chapter 8 Crystallography with X-ray Free Electron Lasers 181
N. A. Zatsepin

8.1 X-ray Free Electron Lasers – An Introduction 181
8.2 Radiation Damage at XFELs 184
8.3 Serial Femtosecond Crystallography 185
8.3.1 SFX Experimental Setup 185
8.3.2 SFX Early Achievements 187
8.3.3 SFX Sample Delivery and Data Collection Rates 189
8.4 Time-resolved Serial Femtosecond Crystallography 192
8.4.1 Pump Probe Serial Femtosecond Crystallography 193
8.4.2 Mix-and-inject Serial Femtosecond Crystallography 194
8.5 Serial Femtosecond Crystallography Data Analysis 194
8.5.1 SFX Data Collection Overview 194
8.5.2 Data Collection Monitoring 195

8.5.3 Hit Finding 195
8.5.4 Bragg Diffraction Analysis: Indexing, Merging, Post-refinement 197
8.5.5 Phasing and Model Refinement 202
8.5.6 *De novo* Phasing of SFX Data 203
8.5.7 SFX Data Volumes and Data Sharing 205
8.6 New Developments 206
8.6.1 Sparse Crystal Pattern Indexing 206
8.6.2 Nanocrystal Shape Transform Phasing 206
8.6.3 Continuous Diffuse Scattering 206
8.6.4 Single-layer 2D Crystals 207
8.6.5 Incoherent Diffractive Imaging 208
8.7 Conclusion 208
Acknowledgements 209
References 209

Subject Index 225

CHAPTER 1

Practical Approaches for In Situ *X-ray Crystallography: from High-throughput Screening to Serial Data Collection*

ISABELLE MARTIEL,[a] VINCENT OLIERIC,[a] MARTIN CAFFREY[b] AND MEITIAN WANG*[a]

[a] Swiss Light Source, Paul Scherrer Institute, CH-5232 Villigen, Switzerland; [b] Membrane Structural and Functional Biology Group, Schools of Medicine and Biochemistry and Immunology, Trinity College, Dublin, Ireland
*Email: meitian.wang@psi.ch

1.1 Introduction

1.1.1 What Exactly Is *In Situ*?

In macromolecular crystallography (MX), *in situ* data collection refers to a diffraction measurement performed on crystals where and as they grow. In other words, the crystals are not harvested individually from their growth environment, as is typically done in standard MX with a harvesting loop. Thus, in the *in situ* experiment, the original growth medium and the

Chemical Biology No. 8
Protein Crystallography: Challenges and Practical Solutions
Edited by Konstantinos Beis and Gwyndaf Evans
© The Royal Society of Chemistry 2018
Published by the Royal Society of Chemistry, www.rsc.org

crystallization compartment remain in place surrounding the crystal during interrogation with the X-ray beam. By contrast, both are removed or minimized in classical loop harvesting protocols to increase diffraction signal-to-noise ratio (SNR) by minimizing background scattering. In the strictest embodiment of an *in situ* experiment, the crystal growth plate or chamber must remain hermetically sealed from the moment the crystallization experiment is set up and data collection must be done at growth temperature. However, many so-called *in situ* measurements are made under conditions departing to varying degrees from this limiting definition.

A few examples illustrate the extent to which the *in situ* label has been used. Jet sample delivery developed at X-ray free electron laser (XFEL) facilities has been considered an *in situ*-like method. In this case, microcrystals remain suspended in the mother liquor or the lipid cubic phase (LCP) where they grew. However, these samples have been transferred between syringes and reservoirs, sometimes filtered, and finally extruded under pressure into an X-ray chamber that is sometimes under vacuum. These post-growth handling steps accompanied by variations in pressure and temperature can mean that data is collected under conditions that are far removed from *in situ*. Several methods, sometimes presented as *in situ* methods, include a mother liquor removal step, such as the Crystal Direct approach[1] (see Section 1.2) and several XFEL solid support sample preparation methods, where the mother liquor is blotted[2] or sucked away[3] to help position crystals into ordered wells. This mother liquor removal distinguishes these preparation methods from *in situ* experiments.

In this chapter, after a general introduction to *in situ* experiments (Section 1.1), we will cover the different *in situ* setups and the evolution of the field, following a historical perspective. *In situ* experiments date back to the period where X-ray capillaries were used to grow crystals by microdialysis and interface diffusion methods, in order to avoid the difficulties of transferring grown crystals into capillaries for data collection.[4] However crystal movement in the capillary often made the technique impractical.[5] In the 1990s, García-Ruiz and coworkers formalized gel-acupuncture methods to collect data on *in situ* counter-diffusion grown crystals in capillaries without any post-growth transfer, at room temperature and under cryogenic conditions.[4,5] In 2004, Jacquamet, Ferrer and coworkers demonstrated the first *in situ* capable automated setup at a synchrotron beamline, where SBS-format crystallization plates were placed in the beam by a robot arm.[6] The automated handling of SBS-format plates has spread in many synchrotron facilities as well as to laboratory X-ray instruments since then (Section 1.2), benefiting in particular the field of virus crystallography.[7] An intense period of development of *in situ*-specific setups started in parallel, towards format reduction, microfluidic and on-chip systems (Section 1.3). The latest phase of development has seen the emergence of *in situ* experiments optimized for serial crystallography and compatible with data collection at cryogenic temperature (Section 1.4).

1.1.2 Goals of *In Situ* Experiments

In situ methods can be used for a variety of purposes at different stages of a project. In the phase of optimizing crystallization conditions, *in situ* screening can help distinguish between protein and salt or small molecule crystals, as a complement to UV fluorescence and second-order harmonic generation techniques.[8] The unique advantage of X-ray screening is the direct access to data collection-relevant information such as diffraction quality, space group and unit cell, which are not provided by optical techniques. *In situ* screening can therefore help to identify genuine protein crystal hits, to find the best diffracting crystal form in the case of polymorphs, or in the search of different space groups,[9] and to diagnose for loss of diffraction quality due to crystal manipulation and/or cryo-cooling. *In situ* screening can help increase the efficiency of the protein-to-structure pipeline by enabling diffraction-based identification of best conditions and ligand binding state. This is especially valuable for drug discovery applications involving ligand screening.[10,11]

In situ experiments are not limited to screening and optimization. In some projects they are used for final data collection and structure solution. This is the case for crystals that cannot be handled with a loop (crystal degradation upon opening of the well or during harvesting) or flash-cooled in liquid nitrogen, *e.g.* in virus crystallography,[7] or for very small crystals, such as virus and *in meso*-grown membrane protein crystals, where harvesting hundreds of crystals for serial crystallography is time-consuming and may not be practical (see Section 1.4). Due to limitations in the tolerable X-ray dose at room temperature and geometrical constraints imposed by some crystallization containers, it is almost impossible to collect a complete data set from a single crystal in certain *in situ* setups, as is usually done in standard cryo-crystallography. Accordingly, partial data sets from several crystals must be combined as practiced in micro- and serial crystallography.[12] Depending on the sample type, data collection can be performed either using a multi-crystal approach or using serial crystallography methods.[13] In the multi-crystal approach, a few partial data sets covering significant angular wedges from a few crystals are merged together. The sorting and merging of data sets are generally performed manually or semi-manually by the crystallographer. In the serial approach, large numbers of small wedges or even still images from many crystals are assembled, which requires automation in data set processing, selection and merging. The serial approach derives from serial femtosecond crystallography (SFX) data collection, where only still images are collected on thousands of randomly oriented small crystals.[14,15] In synchrotron-based serial data collection, wedges of typically a few degrees are collected on each crystal. In both cases, data collection of a complete data set relies on the varied or random orientation of crystals for adequate sampling of reciprocal space. Preferential orientation of the crystals on the plate or well surface is therefore to be minimized or compensated for by tilting the sample support during X-ray data collection.

With *in situ* methods, unnecessary manipulation of crystals by harvesting is avoided. However, harvesting is not always detrimental: clear cases where post-growth treatments such as dehydration increase the diffracting quality have been reported.[16] Methods for controlled dehydration and other post-growth treatments in *in situ* plates have been developed.[17] Another characteristic of manual harvesting is the introduction of a possible source of irreproducibility in the experiment, since two crystals are rarely harvested exactly in the same way, even by the same person. This is less of an issue with *in situ* methods.

Historically, *in situ* measurements are performed mainly at room temperature (RT) (see Section 1.2). RT data collection is often deemed biologically more relevant. Further, it enables the probing of conformational landscapes, time-resolved studies and chemical reactions in the crystals. Measurements at RT usually result in lower crystal mosaicity. In certain cases, such as with virus crystals, RT data collection is the only option due to crystal fragility and sensitivity to cryo-cooling. Recent developments with thin-film samples (see Section 1.4.1) offer the possibility to perform flash-cooling of *in situ* samples and to collect data under cryogenic conditions. Cryo-treatment is not compliant with the strict definition of *in situ*, but low temperature (100 K) data collection has significant advantages that include a 50- to 100-fold increase of the tolerable dose. Further, cryo-cooled samples are easily stored and transported.

1.1.3 Challenges of *In Situ* Methods

Here we list the challenges related to *in situ* experiments, of which users should be aware when selecting a particular method and planning experiments. The first and foremost challenge is the relatively high scattering background arising from the support and the growth medium surrounding the crystal. This generally results in sharp or diffuse scattering rings or arcs at intermediate-to-low scattering angles (∼3–6 Å). Although *in situ* setups are usually optimized to reduce such scatter (see Section 1.1.4), background contribution will remain larger for most *in situ* setups compared to a correctly loop-harvested cryo-cooled crystal.

The second challenge, radiation damage, is not specific to *in situ* experiments. Detecting and managing radiation damage is also crucial for successful data collection with conventional methods.[18] With *in situ* methods, the problem of radiation damage is pronounced when data collection is done at RT and/or with small crystals. At RT, the tolerable dose per crystal is of the order of a fraction of a MGy,[19] while under cryogenic conditions at 100 K a single crystal can take up to about 20 MGy (the so-called Henderson limit[20,21]) for molecular replacement methods, or about 5 MGy for experimental phasing methods.[22] In practice, these should be considered as upper dose limits, since many crystals are more sensitive,[23] in a manner that depends on heavy atom content, crystal composition and crystallization conditions.

The third issue is the geometrical constraints imposed by *in situ* plates and supports, which limit the angular range of data that can be collected. The accessible angles vary with the type of plate and the setup. With some plates it is difficult to accurately position the beam on the crystals due to optical refraction by curved or thick plastic surfaces, and/or on the crystallization drops.[24] For this reason, plates with flat surfaces are preferred especially for small crystals, although they are not always convenient when surface active agents, such as detergents, are present in the crystallization conditions.

Special beamline equipment is necessary to perform most *in situ* measurements. Thus, suitable hardware to transport the plate or support into the beam and bespoke software must be available. Synchrotron facilities often have at least one beamline equipped for *in situ* experiments (Table 1.1). Serial crystallography approaches also require specific data acquisition, processing and merging software to handle the data. It is recommended to process and merge the data online, to monitor and optimize data quality and completeness during data collection.[24]

Since *in situ* methods are at the interface between crystallization and crystallography, several crystallization-related constraints should also be considered in the choice or design of the *in situ* setup. Thinner windows will enable faster evaporation of solvent from solutions inside the plates, such that the drops dry quicker.[25] Overcoming this issue for crystallization experiments that last for weeks requires either a compromise on the film thickness, a double-sandwich type setup to prevent evaporation (see Section 1.4.1), or a humidified plate storage environment. Special lids have also been designed to slow down evaporation.[25] Another point to consider is the compatibility with optical imaging systems (polarized light microscopy, UV fluorescence) to identify crystal hits, and with laboratory liquid handling robotics to set up the drops. Finally, one should bear in mind the influence of interfaces, geometry and drop size on nucleation probabilities and growth processes.[26] Optimization of crystallization conditions for a given type of plate or support is often required.

1.1.4 Enabling Technologies

The success of *in situ* crystallography has been facilitated by the introduction of a number of other technologies. Progress in synchrotron radiation technologies and X-ray optics has led to the introduction of microbeams[12,24] with a high flux density to address ever smaller crystals. The development of fast detectors, such as the PILATUS and EIGER detectors, enabled continuous data collection with weakly diffracting crystals.[34] Specific hardware has been developed to place *in situ* supports in the beam (Section 1.2). Beamline controls and software deliver a high level of automation, first introduced as an integrated setup at beamline BM30-FIP,[35,36] extending now to the fully automated MASSIF-1 beamline at the European Synchrotron Radiation Facility (ESRF).[37] Grid-scan or rastering procedures facilitate localizing

Table 1.1 *In situ* measurement capabilities reported at various synchrotron facilities.

Synchrotron	Beamline	Reported *in situ* capacities	References
APS	SBC 19-ID	SBS: Goniometer	27, https://www.sbc.anl.gov/
APS	GM/CA 23ID-B & D	Thin-film sandwich	28, http://www.gmca.anl.gov/
BESSY II	BL14.1	SBS: MD2 goniometer	29, https://www.helmholtz-berlin.de/forschung/oe/np/gmx/ancillary-facilities/insitu-screening_en.html
DLS	I03	SBS: Goniometer	http://www.diamond.ac.uk/Beamlines/Mx/Equipment-on-Demand/In-situ-Data-Collection.html
DLS	I24	SBS: Horizontal goniometer Thin-film sandwich: on vertical goniometer	http://www.diamond.ac.uk/Beamlines/Mx/Equipment-on-Demand/In-situ-Data-Collection.html
DLS	VMXi	SBS: Goniometer	http://www.diamond.ac.uk/Beamlines/Mx/VMXi.html
ESRF/FIP	BM30	SBS: G-rob robot	6, http://www.fip-bm30a.fr/index.php/trading-hours-and-holidays/manage-diaries/description/10-services-available-on-fip
ESRF/EMBL	ID30B	SBS: Goniometer	http://www.esrf.eu/id30b
ESRF	ID13	LCP jet	30, http://www.esrf.eu/UsersAndScience/Experiments/XNP/ID13
KEK	Several beamlines	SBS: Goniometer	31, http://www2.kek.jp/imss/sbrc/eng/beamline/px.html#beamline
LNLS	W01B-MX2	SBS: G-rob robot	http://lnls.cnpem.br/linhas-de-luz/mx2-en/overview/
MAX IV	BioMax	SBS: ISARA robot	https://www.maxiv.lu.se/accelerators-beamlines/beamlines/biomax/
NSLS II	FMX	SBS: Goniometer	https://www.bnl.gov/ps/beamlines/beamline.php?b=FMX
NSLS II	AMX	SBS: Goniometer	https://www.bnl.gov/ps/beamlines/beamline.php?b=AMX
PETRA III/EMBL	P14	SBS: Goniometer (CrystalDirect plates)	https://www.embl-hamburg.de/services/mx/P14/index.html
SLS	X06DA–PXIII	SBS: CATS robot	9, https://www.psi.ch/sls/pxiii/ https://www.psi.ch/sls/pxiii/crystallisation-facility
SLS	X06SA-PXI & X10SA-PXII	Thin-film sandwich LCP jet	11, 32, https://www.psi.ch/sls/pxi/
SOLEIL	PROXIMA1	SBS: CATS robot	http://www.synchrotron-soleil.fr/Recherche/LignesLumiere/PROXIMA1
Spring-8	BL32XU	Thin-film sandwich	33, https://beamline.harima.riken.jp/en/bl_info/bl32xu_info.html

crystals invisible by optical methods, identifying the best diffracting crystals or regions of crystals and performing diffraction-based crystal centering,[34,38] lately by an automated analysis of the rastering results.[39]

With regard to data processing and management, *in situ* data collection in crystallization plates has been facilitated enormously by powerful multi-crystal merging procedures. Crystal selection[40] and clustering methods[41,42] have also proven useful, as recently reviewed.[43] Data management of the often large number of crystallization trials has also received attention. The recent use of haptic interfaces is one such example.[44]

Early on, great effort was invested in optimizing the materials used to manufacture *in situ* plates and supports. The commercial availability of recently developed low background, UV-friendly specialty polymers, mainly cyclic olefin (co-)polymers, in industrial grades of suitable quality, thickness and affordability, has been integral to the success of the approach.[45] The design of new high-throughput *in situ* consumables often involves materials optimization and polymer processing. The success of a new *in situ* method is also often correlated to the translation into commercially available consumables, and the establishment of user-friendly, easily reproducible protocols.[11,32]

1.2 *In Situ* Screening at the Synchrotron: Standard SBS Plates

1.2.1 Development History

The first demonstration of automated *in situ* experiments on SBS-format plates, by Jean-Luc Ferrer and coworkers at the FIP-BM30A (French beamline for Investigation of Proteins) at the ESRF in 2004, opened a new paradigm for *in situ* experiments, inaugurating the era of high-throughput *in situ* methods. Jacquamet, Ferrer and coworkers[6] developed beamline hardware and software to perform *in situ* data collection directly in the crystallization plate in an easy and efficient manner (Figure 1.1). The SBS format for microplates was established around the year 2000 by the Society for Biomolecular Screening (SBS), now part of the Society for Laboratory Automation and Screening (SLAS), and the American National Standards Institute (ANSI). The goal was to ensure compatibility between plates from different manufacturers and laboratory automation instrumentation for drug discovery research. The 96-well SBS plates are therefore compatible with laboratory robotics used for drop setting in crystallization experiments and are now ubiquitous in crystallization laboratories. Although *in situ* experiments in SBS plates started mainly as a screening tool and are often called '*in situ* plate screening', the collection of complete data sets was performed early on, first on single high symmetry crystals[6] and then by merging data from a small number of crystals.[45] Serial-like data collection of small wedges from a large number of crystals was also later demonstrated.[46]

In situ data collection has been particularly beneficial in structural biology fields where crystals are fragile and difficult to flash-cool. For instance, virus

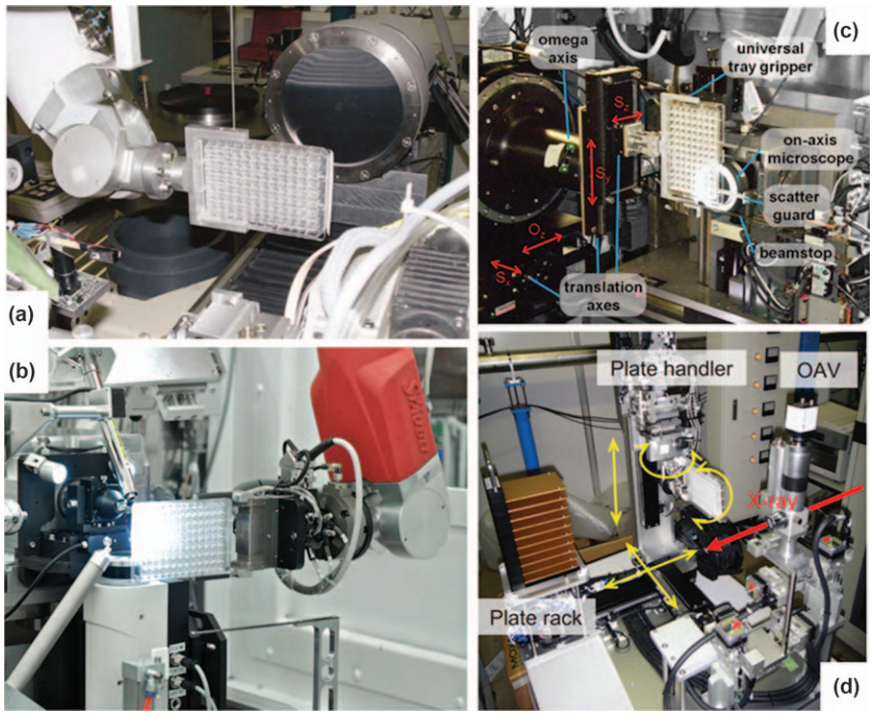


Figure 1.1 (a) The original plate screening setup at beamline FIP-BM30A (ESRF), reproduced from ref. 6. Copyright 2004, with permission from Elsevier.[6] (b) The CATS robot arm in position for data collection on a plate at beamline X06DA-PXIII (SLS), reproduced from ref. 9. Copyright 2011 American Chemical Society. (c) The I24 plate screening goniometer at the Diamond Light Source (DLS), reproduced under a Creative Commons License (https://creativecommons.org/licenses/by/2.0) from Axford *et al.*[24] Copyright 2012, International Union of Crystallography. (d) The PLEX system at the Photon Factory. Reprinted from ref. 31 with the permission of AIP Publishing.

crystals have large unit cells and weak crystal contacts and are notoriously fragile. It is often difficult to find suitable cryo-cooling conditions for such crystals, and the increase of mosaicity often observed upon cryo-cooling can result in spot overlap due to the large unit cell. For these reasons, most virus structures determined by X-ray crystallography are based on data collection at RT.[7,24,47]

Efforts are underway to make *in situ* plate experiments suitable for ligand screening applications. This includes fragment-based screening, which generally involves large numbers of crystals. Well-diffracting crystals are grown under identical crystallization conditions and are soaked (or co-crystallized) using a library of chemical ligands to determine the degree of binding. For fragment-based screening data set collection, high completeness and/or multiplicity is not always required.[45] Ligand addition for *in situ*-like ligand

screening experiments can be performed either using standard liquid handling robots[10] or by acoustic droplet ejection (ADE)[48] using, for example, the commercial system Labcyte Echo 550, or in-house built setups. The very small volumes (down to a few nanoliters) handled by ADE make it possible to multiply the number of crystallization or soaking trials and therefore to screen more ligands.[48] Gelin *et al.*[10] developed a method where the base of each well in the plate is coated with dry ligand. The ligand solubilizes in the dispensed crystallization drop and ideally diffuses into the crystal.

1.2.2 Plate Handling Hardware

Two types of plate handling hardware exist at beamlines for SBS plate *in situ* collection: robots and goniometers. An automatic sample changer robot, normally used to exchange cryo-cooled samples, can be equipped with a special gripper for moving SBS plates to and from a multi-plate hotel. The precision and stability of 6-axis industrial robots, commonly used as sample changers, is sufficient to reliably position the plate, and to center and rotate the crystal as a goniometer would do on the beam axis – by combining the 6-axis degrees of freedom to emulate a single axis rotation distinct from the 6th rotation axis. The robot and beamline control software must also be adapted to enable these complex motions. The precision achieved upon rotation of a well-centered crystal is excellent, as shown by the small beam footprint left on a test crystal after a 60° rotation (figure 4D in Pinker and coworkers[49]). Three examples of this type of system are the commercial CATS,[50] ISARA and G-Rob[45] systems, in use at SLS, BESSY II, Soleil, Max IV and ESRF FIP (Table 1.1). A second approach is to use a dedicated, standard goniometer to move and to rotate the plate. The plate screening goniometer can simply be an adaptor on the main goniometer, or it can be distinct from the main goniometer for single crystal work, in which case fast goniometer switching procedures should be in place. The plate is either fixed manually to the goniometer with an adaptor holder in which the plate is placed or placed by an automatic sample changer. Recent examples (Figure 1.1) of such setups can be found at beamline I24[24,51] of the DLS, or the PLEX system at the Photon Factory.[31] The MD2 diffractometer can also take SBS plates using an adaptor (Table 1.1). SBS plate handling hardware has been developed for laboratory sources.[52] The Rigaku PlateMate system is one such example.

1.2.3 Plate Optimization for *In Situ*

The importance of the material composition and design of the plate for successful *in situ* data collection is fully appreciated.[6] Both the intensity and resolution of the scattering background must be minimized to increase the SNR. In particular, the background around the resolution limit should be minimized to maximize the SNR at the crystal's highest resolution. Amorphous materials are often preferred over crystalline or semi-crystalline materials due to their broader, more diffuse scattering properties.

The optical clarity, low birefringence and UV-compatibility, as well as the fabrication-related properties must also be considered. The intensity of the background generated by scattering from an amorphous material depends on several material- and geometry-related parameters:[53]

$$I_{\mathrm{bg}} \propto A\frac{\rho V}{M_{\mathrm{w}}} f_{\mathrm{bg}}(s)^2 \tag{1.1}$$

where A is an absorption factor (which depends on the absorption coefficient μ_{abs} and thickness of the material), V is the illuminated volume (equal to the product of the beam area and the material thickness), ρ is the material mass density and M_{w} its molecular weight, and f_{bg} is the scattering-angle-dependent structure factor of the material. The proportionality factor, not shown in eqn (1.1), contains factors related to the detector pixel geometry and position, the X-ray beam characteristics, exposure time and physical constants. Upon inspection, eqn (1.1) shows that the background can be reduced by decreasing the material thickness, and selecting materials with suitable absorption and scattering properties, composed preferably of low-Z atoms.

Jacquamet and coworkers[6] compared several of the materials available at the time of their work. They recognized the need for the design of special *in situ* plates, with optimized well geometry and plastic thickness. Such an optimized plate, the Greiner CrystalQuick X45 for sitting-drop experiments, was introduced formally in 2011. The selected material was cyclic olefin copolymer (COC), a specialty plastic with low birefringence properties. The thickness of the well bottom was reduced to 300 μm, and the well shape allowed collection over a total angular range up to 80°. Later a second *in situ* optimized SBS plate was introduced, the MiTeGen InSitu-1 plate,[54] where the drops are directly deposited on a flat COC film of thickness 100 μm. The plates can be used for sitting- or hanging-drop experiments and are compatible with deposition of multiple drops by ADE. It is important to note that water permeability is of concern when using such thin plastic films, meaning that the drops dry faster.[25] In the latest *in situ* plate brought to market, the CrystalDirect plate developed at the European Molecular Biology Laboratory (EMBL) and available from MiTeGen, the COC film thickness has been reduced to 25 μm. The CrystalDirect plate is designed to be compatible with the automated harvesting system of the same name. Here the excess mother liquor is aspirated through a small hole, a pin is glued onto the film, laser photoablation is used to cut the film around the crystals and the glued pin tip and this is followed by immediate flash-cooling.[1,55]

1.2.4 Automation and Pipeline Integration

Automation compatibility was at the heart of the first plate screening experiments in 2004.[6] Further developments logically followed. In 2011, the first integrated plate screening pipeline was established at the Swiss Light

Source (SLS) beamline X06DA-PXIII.[9] In this setup, in addition to a simple short-term plate hotel inside the hutch, the sample changer has direct access to the Formulatrix Rock Imager RI 1000 plate hotel located in the adjacent crystallization facility. A 4-axis robot shuttles the plates through the radiation safety wall between the crystallization facility and beamline hutch. Using this automated system, based on the online access of drop imaging results from the automated imager, users can perform targeted *in situ* diffraction-based screening of their crystallization plates placed in the incubator without any onsite intervention after the setup of the plate. This arrangement facilitates fast feedback on the diffraction quality of the crystals, eliminates the need for risky plate transport or shipping, and enables fully remote plate screening operation. In addition, the X06DA-PXIII setup allows fast (2 minute) exchange by the users between standard cryogenic data collection and *in situ* screening mode.

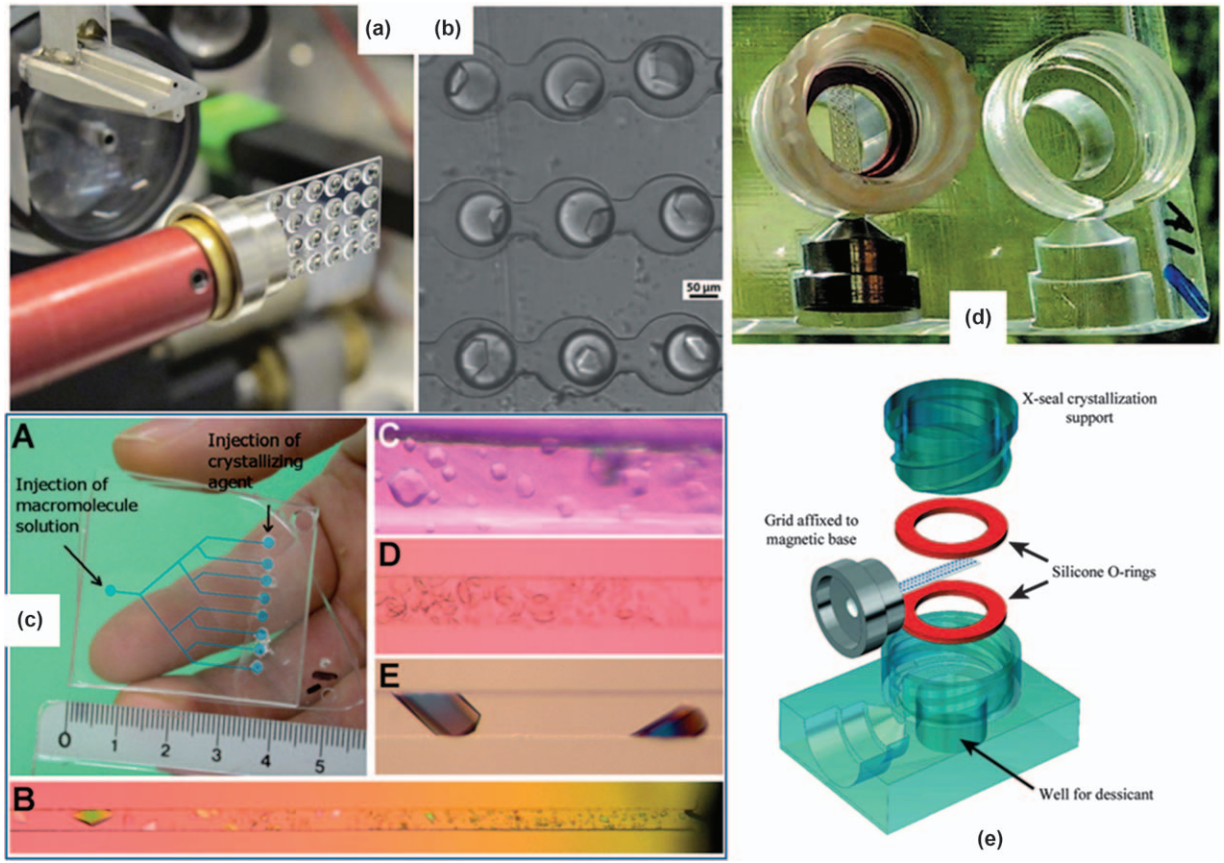
Following this lead, the VMXi beamline,[56] a microfocus beamline dedicated fully to *in situ* plate screening and data collection, has been constructed at DLS. Two Formulatrix Rock Imagers set at two different incubation temperatures are installed in the beamline hutch and are directly accessed by the plate changer robot, which is mounted on a large linear axis. A system for maintaining temperature control on the plate during data collection is also foreseen. This beamline is expected to operate on a fully automated basis, where users mark the positions of interest on the images from the Rock Imager.

1.3 Further Developments: Scale Reduction and Microfluidics

The SBS format had its origins in automated laboratory equipment for liquid handling. In parallel to SBS-format *in situ* developments, there have also been *in situ* developments that use non-SBS formats, better adapted to data collection with standard goniometers at synchrotrons. These setups are often designed for direct data collection rather than screening. The departure from the SBS format goes generally in the direction of a format size reduction, both in the overall footprint, better compatible with crowded sample environments at beamlines, and in the crystallization trial dimensions, as with microfluidic setups. We have here arbitrarily distinguished between small format multi-crystal holders for *in situ* experiments and more classical microfluidics setups.

1.3.1 Small Formats

The X-chip (MiTeGen[61]) developed by Kisselman *et al.*[57] is a small plastic chip with drop positions marked, where micro-batch under oil crystallization trials can be set up (Figure 1.2a,b). Each drop location is defined by concentric hydrophilic–hydrophobic patterned rings, which ensure


(a)
(b)
50 µm
(d)
A
Injection of crystallizing agent
Injection of macromolecule solution
(c)
C
D
E
B
0 1 2 3 4 5
X-seal crystallization support
Grid affixed to magnetic base
Silicone O-rings
Well for dessicant
(e)

good pinning of the water-based crystallization drop and the oil cover layer. The crystallization drop is typically a 50 : 50 by volume mix of protein and precipitant solution (~500 nl total volume), covered with ~1 μl high viscosity oil (paraffin or a paraffin/silicon oil combination, typically). The drops will evaporate after days to weeks, depending on the oil. The chip itself absorbs about 30% of the beam (at 12.4 keV) with its 375 μm thickness and the oil contributing to the absorption and background. Nevertheless, Se-SAD phasing data collection was successfully demonstrated using this setup. To date, all data collection has been done at RT. The chip is fixed on a magnetic base compatible with standard goniometer heads, and the chip dimensions do not exceed this footprint, so that the X-chip is in principle compatible with any standard beamline setup.

More recently, Baxter *et al.*[60] introduced multi-crystal grids compatible with a home-developed tray for *in situ* crystal growth by vapor diffusion (Figure 1.2d,e). The grids consist of a laser-cut polycarbonate sheet of 100 to 200 μm thickness, with an array of holes, backed with a 5 μm polycarbonate foil. The holes are 125 to 400 μm in diameter. The grids are fixed on standard magnetic bases. In the *in situ* setup, the grid holes are filled with the protein solution and precipitant mixture, either with a liquid handling robot or by ADE. The grids are then installed in the vapor diffusion chamber, sealed with rubber O-rings and a removable lid. The chamber is opened after crystal growth. This type of multi-crystal mount is suitable for goniometer-based data collection both at synchrotrons and XFELs.[62]

ADE-assisted preparation of *in situ* samples has the potential for ligand or fragment screening experiments, as demonstrated by Yin and coworkers with *in situ* experiments set up on micromeshes.[63] Previously, Berger and coworkers[64] have shown that it is possible to grow crystals directly in a loop and to cryo-cool them.

1.3.2 Microfluidic Methods for *In Situ*

Microfluidics is the technique of choice for manipulating small volumes of liquids in a controlled manner. These *in situ* setups offer both the possibility to screen for various crystallization conditions and to collect diffraction data.

Figure 1.2 (a) X-CHIP with 24 wells mounted on a goniometer, reproduced under a Creative Commons License (https://creativecommons.org/licenses/by/2.0) from Kisselman *et al.*[57] Copyright © Kisselman *et al.* 2011. (b) Droplet-based microfluidic device for Laue diffraction on *in situ* grown glucose isomerase crystals, reproduced under a Creative Commons License (https://creativecommons.org/licenses/by/2.0) from Heymann *et al.*[58] Copyright © Michael Heymann *et al.* 2014. (c) On-chip counter diffusion chip (A), and channels with crystals of thaumatin (B), bovine insulin (C), a plant virus (D) and turkey egg-white lysozyme (E). Reproduced from Dhouib *et al.*[59] with permission from The Royal Society of Chemistry. (d, e) High density multi-crystal grids with *in situ* tray, reproduced under a Creative Commons License (https://creativecommons.org/licenses/by/2.0) from Baxter *et al.*[60]

Three types of on-chip crystallization experiments with *in situ* diffraction capabilities can be distinguished: free interface diffusion (FID), counter-diffusion and droplet-based batch. Most devices designed for on-chip data collection use COC as an X-ray-friendly material. However, new materials such as graphene have been tested and are of interest for their water-impermeability and ultralow-background properties.[65]

The main commercial option for FID microfluidics experiments is the Topaz chip in SBS format by Fluidigm,[66] which has been reported to be diffraction-compatible.[9] The Topaz system relies on the use of pressure-activated valves which bring into contact the preloaded precipitant and protein solutions, in up to 96 different conditions. FID experiments are characterized by small reaction chambers in which equilibration by diffusion is achieved relatively quickly and without convective mixing. As a result, the crystals produced by FID are potentially better ordered, and the trajectory in the crystallization phase diagram is better controlled compared to batch experiments.[67] Multilayer valve-based microfluidic devices optimized for *in situ* diffraction have also been reported,[68] with applications in Laue diffraction[69] and for *in meso* crystal growth.[70] Microfluidic FID experiments can screen conditions using small volumes, but the devices are usually difficult or expensive to fabricate and require a pump to operate.

Counter-diffusion differs from FID by the establishment of a gradient of conditions, by diffusion of chemical species over larger distances than in FID. In a single experiment a continuum of crystallization conditions is probed. Counter-diffusion in capillaries was among the first *in situ* diffraction setups,[4,5] and microfluidics soon appeared as a natural scale-down option, while offering more flexibility for channel design. Two groups, Ng and coworkers[71] and Dhouib and coworkers,[59] developed in parallel *in situ* counter-diffusion microfluidic chips. The device by Ng and coworkers[71] consists of single channels, and is commercialized by Greiner BioOne under the name CrystalSlide. Four CrystalSlides can be presented to the beam in a special SBS-format holder. In the commercial version, individual channels can be separated and mounted on a magnetic base.[72] The device by Dhouib and coworkers,[59] as well as the ChipX by Pinker and coworkers,[49] offers the possibility to screen different precipitant formulations against a single protein solution *via* channel branching (Figure 1.2c). The CrystalHarp system is an array of polyimide-coated quartz capillaries presented in SBS format,[9] commercialized by Molecular Dimensions.[73] Counter-diffusion devices are generally filled using pipettes, thus not requiring pump equipment.

Droplet-based microfluidics crystallization experiments are essentially microbatch-under-oil experiments. Each nanoliter trial droplet is separated from the others by a continuum of fluorocarbon oil. The droplets are produced by mixing two or more aqueous solutions, typically protein solution, buffer and precipitant, at the junction where the water-in-oil emulsion is created. The droplets are then stored on the device. Pumping equipment and careful flow or pressure control are required to create the droplets and to vary

the crystallization conditions. In the initial *in situ* droplet-based microfluidics measurements,[74] droplets were produced in devices made from PDMS, and stored for data collection in a 180 μm inner diameter glass capillary coupled to the device. For *in situ* data collection, the capillary containing the droplets was cut and sealed, and fixed on a magnetic base. The commercial CrystalCard device,[75,76] by Protein BioSolutions, works on the same principle. The crystals produced can be harvested or measured *in situ*, either directly inside the chip or by coupling with a capillary.[77] The Plug Maker system includes the pumping equipment and automated controls needed to use the CrystalCard devices. More recently, other X-ray-friendly chips for droplet-based *in situ* experiments have been designed by Heymann *et al.*,[58] using thin COC films for device fabrication. The suitability of the device for serial Laue diffraction data collection at RT was demonstrated. The effects of a confined droplet environment on nucleation and crystallization processes were studied in detail.[78] It was found that a preliminary screening step makes it possible to find conditions where only a single crystal per droplet is obtained, which is an optimal situation for data collection. This was attributed to a confinement-induced negative feedback on the nucleation probability after the first nucleus appeared.

1.4 The Emergence of Serial *In Situ* Data Collection

1.4.1 Thin-film Sandwiches

In recent years, a new class of *in situ* setups has been developed, that we will refer to here as thin-film sandwiches to distinguish them from the previously described *in situ* setups. The motivation for these new developments is to offer a user-friendly setup that can be prepared with standard crystallization equipment and that is compatible with *in situ* serial crystallography. The principle of thin-film sandwich setups is to perform the crystallization trial in a confined space between two thin, X-ray compatible films. To avoid dehydration caused by water permeability of the film, the sandwich is enclosed in a second thick glass or plastic sandwich for the duration of the crystallization experiment and this is removed just before data collection (Figure 1.3). These methods are appropriately called double-sandwich methods. The thin-film sandwich plate has an SBS 96-well plate format that is compatible with laboratory drop setting robotics. In contrast to SBS *in situ* plates (Section 1.2), individual wells can be easily removed from the plate. This allows for direct mounting of wells on standard goniometers. Importantly, individual wells can be flash-cooled in liquid nitrogen, which extends *in situ* crystallography from mainly a screening technique at RT to a routine data collection method at cryogenic temperature. The wells can be fixed on standard pins, flash-cooled, shipped in a dry-shipper and mounted on a goniometer with an automated sample changer as commonly practiced in single crystal cryo-crystallography. Therefore, beam interrogation on *in situ* thin-film sandwiches can be performed either at RT[11] or under cryogenic

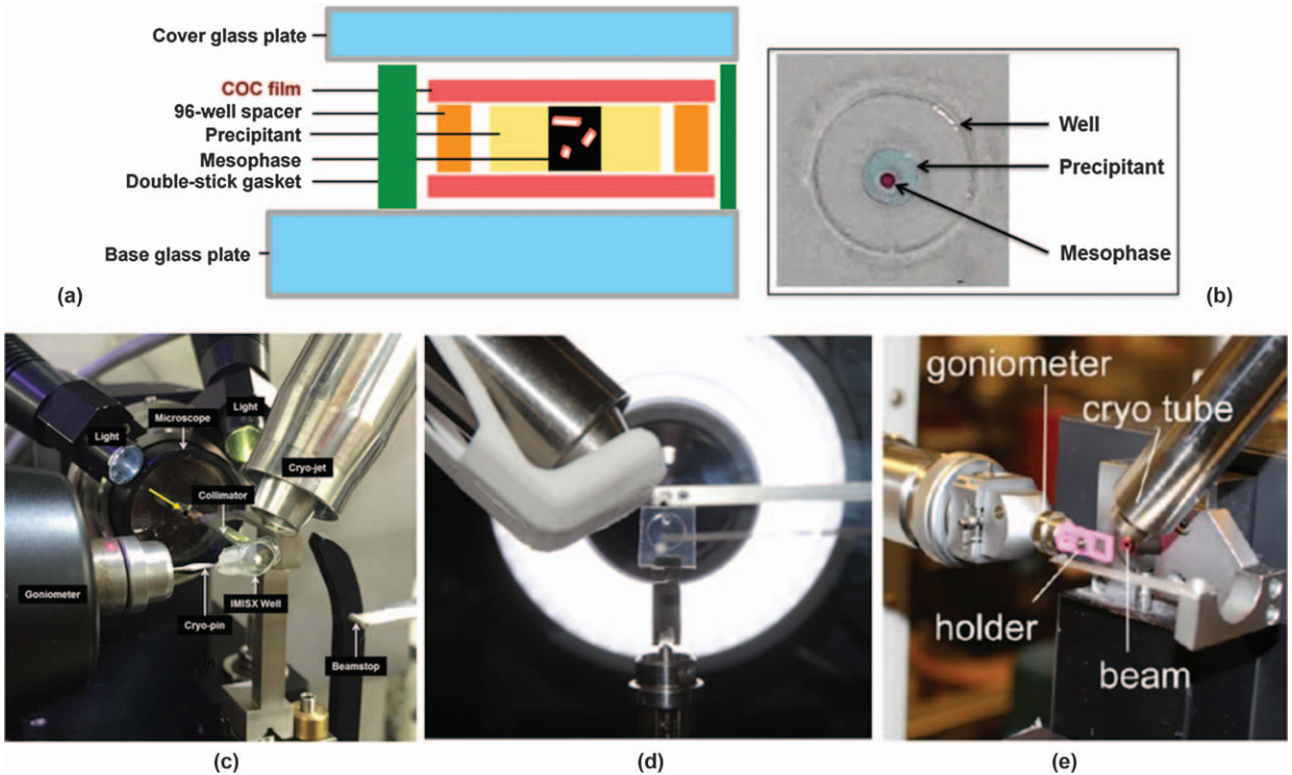


Figure 1.3 Schematic (a) and picture (b) of a well of the IMISX plate, reproduced under a Creative Commons License (https://creativecommons.org/licenses/by/2.0) from Huang *et al.*[11] (c) Cryo-cooled COC IMISX well at the X06SA-PXI beamline at the SLS, reproduced under a Creative Commons License (https://creativecommons.org/licenses/by/2.0) from Huang *et al.*[32] (d) Room temperature COP double sandwich setup at DLS I24, reproduced under a Creative Commons License (https://creativecommons.org/licenses/by/2.0) from Axford *et al.*[79] (e) Mylar double sandwich setup at the APS GM/*CA* beamlines, reproduced from ref. 28 (http://pubs.acs.org/doi/abs/10.1021/acs.cgd.6b00950), with permission from the American Chemical Society.[28]

conditions.[32] The flat geometry also offers a potentially larger data collection angular range compared to SBS *in situ* plates with curved wells, which is particularly attractive for application in the emerging serial crystallography field. Clearly-explained procedures for plate setup and easy-to-handle commercial solutions are now available,[11,32] providing thin-film sandwich methods with opportunities for rapid expansion and wide-spread use.

Careful selection of the thin film material is necessary to minimize the absorption and diffraction background contribution. A material with negligible absorption and background scattering would be ideal. However, in practice, it is sufficient if the absorption and background scattering of the thin film is low in comparison to the contribution from the crystallization medium. In addition, particular care must be taken to minimize the background scattering near the diffraction resolution limit of the crystals, where the diffraction signals are weak. This is usually around 2.5–4 Å for most targets. Plastic films are often used for their low cost, easy handling and the commercial availability of films with relatively low thicknesses. Nonetheless, they can have ring-featured background[28] and their water-tightness is often relatively low. Reported film variants include 25 μm COC,[11] 13 μm cyclic olefin polymer (COP),[79] 8 μm Kapton[80] and 3.5 μm Mylar.[28] The currently available commercial setups use 25 μm COC (IMISX™ by MiTeGen[81]) and 40 μm plastic film (DiffraX™ by Molecular Dimensions[82]). Other materials such as silicon nitride membranes, with sub-micrometer to nanometer thicknesses and low water permeability, have also been used. However, these are fragile, difficult to handle, are more expensive,[83] and are used mainly for data collection purposes. The thickness of the spacer between the two thin films in the sandwich defines sample thickness. The spacer thickness is therefore a major parameter in the optimization of background. The spacer thicknesses reported in the literature range from 50 to 140 μm. Commercial setups come with 140 μm (IMISX™ by MiTeGen[81]) and 100 μm (DiffraX™ by Molecular Dimensions[82]) spacers. In the DiffraX™ setup by Molecular Dimensions, the spacer is already fixed on the base film for ease of handling. Thinner spacers are commercially available. However, issues of preferential orientation of crystals, influence on the crystallization conditions and difficulty of handling have been reported with thinner spacers.[28] Figure 1.4 shows representative background curves corresponding to the contribution of each of the components of the thin-film sandwich in the current IMISX™ setup by MiTeGen. The COC film has maximum scatter at intermediate resolution (4–6 Å), while at higher scattering angles the matrix (LCP and precipitant solution) gives the most significant contribution because of the spacer thickness. Thin silicon nitride has virtually zero background, which becomes beneficial compared to plastic films in cases where the spacer used is relatively thin. Dedicated holders for securing the sample on standard magnetic goniometers have been developed by several groups[28,79] and some are commercially available.

Thin-film methods were developed originally for *in meso* or LCP crystallization, since the high viscosity of the mesophase in which crystallization takes place makes it difficult to harvest crystals, and the method often yields

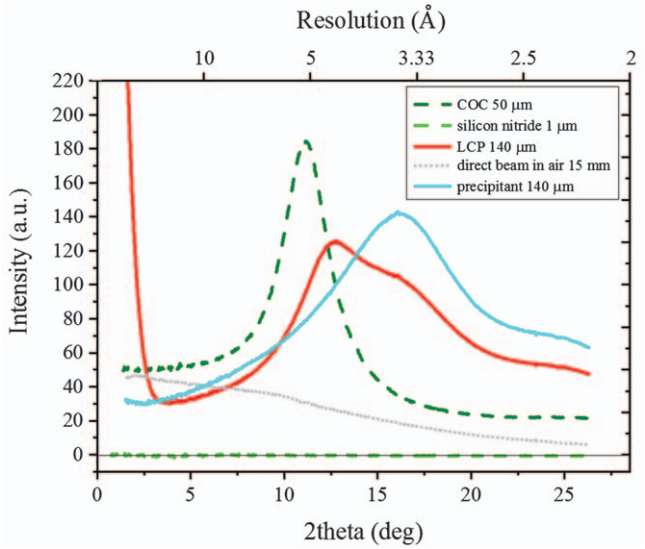


Figure 1.4 Representative background contribution curves from the different components of a typical IMISX plate: 2×25 µm COC, 140 µm LCP or precipitant matrix (from the spacer thickness). The contribution curves were obtained by deconvolution from total background data taken at various points of the well. For comparison, the contribution of 1 µm silicon nitride is also shown, as well as the scattering from 15 mm of direct beam path in air (obtained by subtraction from images taken with two different beamstop distances). The data was measured at beamline X06SA-PXI of the SLS, at 12.67 keV, 1 s exposure with flux of 4×10^{11} ph s^{-1}, beam size 20 µm×10 µm, detector distance 400 mm, and beamstop distances 10 and 25 mm.

small crystals. The sandwich film, of controlled thickness and flat geometry, provides a conveniently rigid scaffold with which to handle the sample, and to clearly view the crystals just as with standard glass LCP plates. On the other hand, thanks to the mesophase viscosity, handling of the samples is possible without perturbing the crystal growth environment. It has been demonstrated that thin-film sandwich setups are compatible with crystallization and data collection for water soluble proteins as well, with and without a mesophase growth medium.[28,79]

Data collection and screening are typically performed in the same way in thin-film sandwich setups. Small wedges of data are collected on a large number of micro-crystals in serial fashion at a microfocus beamline. This mode of data collection is very well suited for full automation in combination with rastering. Examples of such automated serial collection utilities include the MeshAndCollect system[39] at the ESRF, the Zoo system[84] at SPring-8 and the CY+ system at SLS (unpublished).

Thin-film sandwich setups have been demonstrated by Huang *et al.*[11,32] to be compatible with data collection for experimental phasing. This includes

bromine and native SAD phasing of various proteins. Key to the success of the process was the accumulation of enough data to extract weak anomalous signals. Schubert *et al.*[80] have explored the suitability of the setup for time-resolved dynamic studies at RT, using a dose-dependent study of the progress of radiation damage on a model protein as an example. The number of crystals required for a complete data set depends on the conditions (RT or cryo), crystal size, space group and phasing method. For example, in the work of Huang *et al.*, in the case of lysozyme at RT,[11] about 100 crystals of around 20 μm in size from 2 wells were needed to solve the structure by molecular replacement, 200 crystals from 4 wells were needed for bromide single wavelength anomalous diffraction (SAD) phasing, while 1000 crystals of from 12 wells were needed for native SAD phasing. Under cryogenic conditions,[32] only a handful of crystals were needed in similar circumstances. For instance, only six 30 μm crystals were required to solve an insulin structure by native SAD. For membrane proteins, significantly more crystals are generally required due to their smaller size, weaker diffraction and enhanced radiation sensitivity. Typically, with a microfocused beam, a few images per crystal can be obtained at RT, and a partial data set under cryogenic conditions, depending on the radiation damage threshold.

1.4.2 Liquid Manipulation Methods

The liquid manipulation methods briefly covered in this section might be considered *in situ* by the absence of manual crystal handing. Post-crystal growth, ADE methods are an emerging sample delivery scheme. They come with a few variants, but all involve the use of acoustic waves of defined frequency propagating through a liquid suspension to deform the surface so as to create droplets of controllable size. Crystals can be trapped in the droplets, which are either presented directly to the beam, ideally in a drop-on-demand fashion,[85] or are deposited on a conveyor belt or tape drive.[86,87,98] Another variant involves trapping the drop in an acoustic standing wave field.[88] However, acoustic droplet manipulation can be difficult in the presence of surfactants, as it is often the case for crystallization of membrane proteins in solution.

Inherited from XFEL sample delivery techniques, injection methods can be compared to *in situ* methods, at least when the crystals are not filtered, pressurized or transferred to or mixed with a different matrix after crystal growth.[89] This corresponds to cases of microcrystals grown in liquid by batch methods and directly injected in a capillary[90] or in a microfluidic trap.[91] Electrospinning injection[92,93] is another liquid delivery technique where mixing is not required. Crystals grown in LCP can also be injected sufficiently slowly for synchrotron serial data collection[30,94,95] using a high-viscosity injector. Injection delivery methods have been covered by several previous reviews,[96] to which the interested reader is referred. Manipulation of microcrystals often involves pipetting, which can be considered as relatively mild handling compared to standard harvesting.[97]

1.5 Conclusion and Outlook

In this chapter we have covered the wide variety of *in situ* crystal growth and diffraction setups available, including SBS-format plates, microfluidics, and thin-film sandwich methods. *In situ* method development is a dynamic field, where new approaches, materials and equipment are being introduced by user groups and facilities on a regular basis. All the methods covered here will continue to benefit from progress in materials manufacturing, for further optimization of thickness and background properties of plates and films.

The development of synchrotron sources will give access to increasingly higher flux densities with the emergence of diffraction limited storage rings at 4th generation synchrotrons. These include the new MAX IV facility in Sweden and the planned upgrades at many 3rd generation sources. The low emittance of this new type of facility naturally increases the flux density and makes it easier to obtain stable microfocused beams useful for *in situ* data collection. Also on the horizon are 'pink beam' beamlines. These provide bandwidths of the order of 0.1–1% *via* multilayer monochromator, in contrast to silicon (111) crystal monochromators with a bandwidth of ~0.02%. The wider bandwidth results in an increase of flux density but also broadens reflections and increases scattering background, which might lower the SNR of weak reflections and create problems due to reflection overlap when used with large unit cell crystals.

Future developments in *in situ* data collection will aim to optimize SNR for smaller crystals and to improve experimental phasing possibilities, in particular for thin-film sandwich setups. *In situ* experimental phasing using heavy atom derivatives and native lighter anomalous scatterers (sulfur, phosphorous, calcium, *etc.*) has already been demonstrated. However, measurement of very small anomalous differences still requires large amounts of data and careful optimization of the SNR, and radiation damage remains an issue. Improvements in crystallization setups and materials, beamline automation and data processing will contribute to making serial experimental phasing a more routine data collection method. One of the next avenues to explore will be the use of serial *in situ* data collection for ligand screening and fragment based drug design. This type of high-throughput application will require improved automation of data collection and data processing. Finally, serial *in situ* techniques call for specific user training to make the new techniques available to all. Towards this end, detailed protocols have been published, including instructional videos,[11,32] and training workshops take place regularly in different facilities.

Acknowledgements

We acknowledge Laura Vera, May Marsh and Chia-Ying Huang for stimulating discussions.

References

1. U. Zander, G. Hoffmann, I. Cornaciu, J.-P. Marquette, G. Papp, C. Landret, G. Seroul, J. Sinoir, M. Röwer, F. Felisaz, S. Rodriguez-Puente, V. Mariaule, P. Murphy, M. Mathieu, F. Cipriani and J. A. Márquez, *Acta Crystallogr., Sect. D: Struct. Biol.*, 2016, **72**, 454–466.
2. P. Roedig, I. Vartiainen, R. Duman, S. Panneerselvam, N. Stübe, O. Lorbeer, M. Warmer, G. Sutton, D. I. Stuart, E. Weckert, C. David, A. Wagner and A. Meents, *Sci. Rep.*, 2015, **5**, 10451.
3. C. Mueller, A. Marx, S. W. Epp, Y. Zhong, A. Kuo, A. R. Balo, J. Soman, F. Schotte, H. T. Lemke, R. L. Owen, E. F. Pai, A. R. Pearson, J. S. Olson, P. A. Anfinrud, O. P. Ernst and R. J. Dwayne Miller, *Struct. Dyn.*, 2015, **2**, 054302.
4. J. M. Garciaruiz, A. Moreno, C. Viedma and M. Coll, *Mater. Res. Bull.*, 1993, **28**, 541–546.
5. F. J. López-Jaramillo, J. M. García-Ruiz, J. A. Gavira and F. Otálora, *J. Appl. Crystallogr.*, 2001, **34**, 365–370.
6. L. Jacquamet, J. Ohana, J. Joly, F. Borel, M. Pirocchi, P. Charrault, A. Bertoni, P. Israel-Gouy, P. Carpentier, F. Kozielski, D. Blot and J.-L. Ferrer, *Structure*, 2004, **12**, 1219–1225.
7. C. Porta, A. Kotecha, A. Burman, T. Jackson, J. Ren, S. Loureiro, I. M. Jones, E. E. Fry, D. I. Stuart and B. Charleston, *PLoS Pathog.*, 2013, **9**, e1003255.
8. J. A. Newman, S. Zhang, S. Z. Sullivan, X. Y. Dow, M. Becker, M. J. Sheedlo, S. Stepanov, M. S. Carlsen, R. M. Everly, C. Das, R. F. Fischetti and G. J. Simpson, *J. Synchrotron Radiat.*, 2016, **23**, 959–965.
9. R. Bingel-Erlenmeyer, V. Olieric, J. P. A. Grimshaw, J. Gabadinho, X. Wang, S. G. Ebner, A. Isenegger, R. Schneider, J. Schneider, W. Glettig, C. Pradervand, E. H. Panepucci, T. Tomizaki, M. Wang and C. Schulze-Briese, *Cryst. Growth Des.*, 2011, **11**, 916–923.
10. M. Gelin, V. Delfosse, F. Allemand, F. Hoh, Y. Sallaz-Damaz, M. Pirocchi, W. Bourguet, J. L. Ferrer, G. Labesse and J. F. C. C. Guichou, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2015, **71**, 1777–1787.
11. C. Y. Huang, V. Olieric, P. Ma, E. Panepucci, K. Diederichs, M. Wang and M. Caffrey, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2015, **71**, 1238–1256.
12. J. L. Smith, R. F. Fischetti and M. Yamamoto, *Curr. Opin. Struct. Biol.*, 2012, **22**, 602–612.
13. K. Diederichs and M. Wang, in *Protein Crystallography: Methods and Protocols*, 2017, ch. 10.
14. I. Schlichting, *IUCrJ*, 2015, **2**, 246–255.
15. J. M. Martin-Garcia, C. E. Conrad, J. Coe, S. Roy-Chowdhury and P. Fromme, *Arch. Biochem. Biophys.*, 2016, **602**, 32–47.
16. C. Abergel, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2004, **60**, 1413–1416.
17. A. Douangamath, P. Aller, P. Lukacik, J. Sanchez-Weatherby, I. Moraes and J. Brandao-Neto, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2013, **69**, 920–923.

18. E. F. Garman, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2010, **66**, 339–351.
19. R. J. Southworth-Davies, M. A. Medina, I. Carmichael and E. F. Garman, *Structure*, 2007, **15**, 1531–1541.
20. R. Henderson, *Proc. R. Soc. B*, 1990, **241**, 6–8.
21. R. L. Owen, E. Rudino-Pinera and E. F. Garman, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 4912–4917.
22. T. Weinert, V. Olieric, S. Waltersperger, E. Panepucci, L. Chen, H. Zhang, D. Zhou, J. Rose, A. Ebihara, S. Kuramitsu, D. Li, N. Howe, G. Schnapp, A. Pautsch, K. Bargsten, A. E. Prota, P. Surana, J. Kottur, D. T. Nair, F. Basilico, V. Cecatiello, S. Pasqualato, A. Boland, O. Weichenrieder, B. C. Wang, M. O. Steinmetz, M. Caffrey and M. Wang, *Nat. Methods*, 2015, **12**, 131–133.
23. J. M. Holton, *J. Synchrotron Radiat.*, 2009, **16**, 133–142.
24. D. Axford, R. L. Owen, J. Aishima, J. Foadi, A. W. Morgan, J. I. Robinson, J. E. Nettleship, R. J. Owens, I. Moraes, E. E. Fry, J. M. Grimes, K. Harlos, A. Kotecha, J. S. Ren, G. Sutton, T. S. Walter, D. I. Stuart and G. Evans, *Acta Crystallogr., Sect. D: Struct. Biol.*, 2012, **68**, 592–600.
25. L. E. Zipper, X. Aristide, D. P. Bishop, I. Joshi, J. Kharzeev, K. B. Patel, B. M. Santiago, K. Joshi, K. Dorsinvil, R. M. Sweet and A. S. Soares, *Acta Crystallogr., Sect. F: Struct. Biol. Commun.*, 2014, **70**, 1707–1713.
26. V. Cherezov and M. Caffrey, *J. Appl. Crystallogr.*, 2006, **39**, 604–606.
27. K. Michalska, K. Tan, C. Chang, H. Li, C. Hatzos-Skintges, M. Molitsky, R. Alkire and A. Joachimiak, *J. Synchrotron Radiat.*, 2015, **22**, 1386–1395.
28. J. Broecker, V. Klingel, W.-L. Ou, A. R. Balo, D. J. Kissick, C. M. Ogata, A. Kuo and O. P. Ernst, *Cryst. Growth Des.*, 2016, **16**, 6318–6326.
29. U. Mueller, N. Darowski, M. R. Fuchs, R. Forster, M. Hellmig, K. S. Paithankar, S. Puhringer, M. Steffien, G. Zocher and M. S. Weiss, *J. Synchrotron Radiat.*, 2012, **19**, 442–449.
30. P. Nogly, D. James, D. Wang, T. A. White, N. Zatsepin, A. Shilova, G. Nelson, H. Liu, L. Johansson, M. Heymann, K. Jaeger, M. Metz, C. Wickstrand, W. Wu, P. Båth, P. Berntsen, D. Oberthuer, V. Panneels, V. Cherezov, H. N. Chapman, G. Schertler, R. Neutze, J. Spence, I. Moraes, M. Burghammer, J. Standfuss and U. Weierstall, *IUCrJ*, 2015, **2**, 1–9.
31. Y. Yamada, M. Hiraki, N. Matsugaki, R. Kato and T. Senda, Proceedings of the 12th International Conference on Synchrotron Radiation Instrumentation – SRI2015, New York, USA, 2015.
32. C.-Y. Huang, V. Olieric, P. Ma, N. Howe, L. Vogeley, X. Liu, R. Warshamanage, T. Weinert, E. Panepucci, B. Kobilka, K. Diederichs, M. Wang and M. Caffrey, *Acta Crystallogr., Sect. D: Struct. Biol.*, 2016, **72**, 93–112.
33. K. Hirata, Y. Kawano, G. Ueno, K. Hashimoto, H. Murakami, K. Hasegawa, T. Hikima, T. Kumasaka and M. Yamamoto, *J. Phys.: Conf. Ser.*, 2013, **425**, 012002.

34. J. A. Wojdyla, E. Panepucci, I. Martiel, S. Ebner, C.-Y. Huang, M. Caffrey, O. Bunk and M. Wang, *J. Appl. Crystallogr.*, 2016, **49**, 944–952.
35. L. Jacquamet, J. Ohana, J. Joly, P. Legrand, R. Kahn, F. Borel, M. Pirocchi, P. Charrault, P. Carpentier and J. L. Ferrer, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2004, **60**, 888–894.
36. M. Roth, P. Carpentier, O. Kaïkati, J. Joly, P. Charrault, M. Pirocchi, R. Kahn, E. Fanchon, L. Jacquamet, F. Borel, A. Bertoni, P. Israel-Gouy and J.-L. Ferrer, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2002, **58**, 805–814.
37. D. Nurizzo, M. W. Bowler, H. Caserotto, F. Dobias, T. Giraud, J. Surr, N. Guichard, G. Papp, M. Guijarro, C. Mueller-Dieckmann, D. Flot, S. McSweeney, F. Cipriani, P. Theveneau and G. A. Leonard, *Acta Crystallogr., Sect. D: Struct. Biol.*, 2016, **72**, 966–975.
38. V. Cherezov, M. A. Hanson, M. T. Griffith, M. C. Hilgart, R. Sanishvili, V. Nagarajan, S. Stepanov, R. F. Fischetti, P. Kuhn and R. C. Stevens, *J. R .Soc., Interface*, 2009, **6**(Suppl 5), S587–S597.
39. U. Zander, G. Bourenkov, A. N. Popov, D. de Sanctis, O. Svensson, A. A. McCarthy, E. Round, V. Gordeliy, C. Mueller-Dieckmann and G. A. Leonard, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2015, **71**, 2328–2343.
40. G. Assmann, W. Brehm and K. Diederichs, *J. Appl. Crystallogr.*, 2016, **49**, 1021–1028.
41. R. Giordano, R. M. Leal, G. P. Bourenkov, S. McSweeney and A. N. Popov, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2012, **68**, 649–658.
42. J. Foadi, P. Aller, Y. Alguel, A. Cameron, D. Axford, R. L. Owen, W. Armour, D. G. Waterman, S. Iwata and G. Evans, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2013, **69**, 1617–1632.
43. P. Aller, J. Sanchez-Weatherby, J. Foadi, G. Winter, C. M. C. Lobley, D. Axford, A. W. Ashton, D. Bellini, J. Brandao-Neto, S. Culurgioni, A. Douangamath, R. Duman, G. Evans, S. Fisher, R. Flaig, D. R. Hall, P. Lukacik, M. Mazzorana, K. E. McAuley, V. Mykhaylyk, R. L. Owen, N. G. Paterson, P. Romano, J. Sandy, T. Sorensen, F. von Delft, A. Wagner, A. Warren, M. Williams, D. I. Stuart and M. A. Walsh, *Methods Mol. Biol.*, 2015, **1261**, 233–253.
44. A. E. Bruno, A. S. Soares, R. L. Owen and E. H. Snell, *J. Appl. Crystallogr.*, 2016, **49**, 2082–2090.
45. A. le Maire, M. Gelin, S. Pochet, F. Hoh, M. Pirocchi, J. F. Guichou, J. L. Ferrer and G. Labesse, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2011, **67**, 747–755.
46. D. Axford, J. Foadi, N.-J. Hu, H. G. Choudhury, S. Iwata, K. Beis, G. Evans and Y. Alguel, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2015, **71**, 1228–1237.
47. A. Burkhardt, A. Wagner, M. Warmer, R. Reimer, H. Hohenberg, J. Ren, E. E. Fry, D. I. Stuart and A. Meents, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2013, **69**, 308–312.

48. E. Teplitsky, K. Joshi, D. L. Ericson, A. Scalia, J. D. Mullen, R. M. Sweet and A. S. Soares, *J. Struct. Biol.*, 2015, **191**, 49–58.
49. F. Pinker, M. Brun, P. Morin, A.-L. Deman, J.-F. Chateaux, V. Oliéric, C. Stirnimann, B. Lorber, N. Terrier, R. Ferrigno and C. Sauter, *Cryst. Growth Des.*, 2013, **13**, 3333–3340.
50. L. Jacquamet, J. Joly, A. Bertoni, P. Charrault, M. Pirocchi, X. Vernede, F. Bouis, F. Borel, J. P. Périn, T. Denis, J. L. Rechatin and J. L. Ferrer, *J. Synchrotron Radiat.*, 2009, **16**, 14–21.
51. V. Grama, D. Axford, G. Duller, M. Burt and R. L. Owen, I24 Endstation upgrade - overview and engineering design, MEDSI 2014 Conference Melbourne, Australia, 2014.
52. D. Hargreaves, *J. Appl. Crystallogr.*, 2012, **45**, 138–140.
53. J. M. Holton, S. Classen, K. A. Frankel and J. A. Tainer, *FEBS J.*, 2014, **281**, 4046–4060.
54. MiTeGen, In-Situ-01™Crystallization Plate, http://www.mitegen.com/products/plates/insitu1/brochure.pdf (accessed on 03.02.2017).
55. F. Cipriani, M. Röwer, C. Landret, U. Zander, F. Felisaz and J. A. Márquez, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2012, **68**, 1393–1399.
56. VMXi beamline at Diamond Light Source, http://www.diamond.ac.uk/Beamlines/Mx/VMXi.html (accessed on 19.02.2017).
57. G. Kisselman, W. Qiu, V. Romanov, C. M. Thompson, R. Lam, K. P. Battaile, E. F. Pai and N. Y. Chirgadze, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2011, **67**, 533–539.
58. M. Heymann, A. Opathalage, J. L. Wierman, S. Akella, D. M. E. Szebenyi, S. M. Gruner and S. Fraden, *IUCrJ*, 2014, **1**, 349–360.
59. K. Dhouib, C. Khan Malek, W. Pfleging, B. Gauthier-Manuel, R. Duffait, G. Thuillier, R. Ferrigno, L. Jacquamet, J. Ohana, J.-L. Ferrer, A. Théobald-Dietrich, R. Giegé, B. Lorber and C. Sauter, *Lab Chip*, 2009, **9**, 1412–1421.
60. E. L. Baxter, L. Aguila, R. Alonso-Mori, C. O. Barnes, C. A. Bonagura, W. Brehmer, A. T. Brunger, G. Calero, T. T. Caradoc-Davies, R. Chatterjee, W. F. Degrado, J. S. Fraser, M. Ibrahim, J. Kern, B. K. Kobilka, A. C. Kruse, K. M. Larsson, H. T. Lemke, A. Y. Lyubimov, A. Manglik, S. E. McPhillips, E. Norgren, S. S. Pang, S. M. Soltis, J. Song, J. Thomaston, Y. Tsai, W. I. Weis, R. A. Woldeyes, V. Yachandra, J. Yano, A. Zouni and A. E. Cohen, *Acta Crystallogr., Sect. D: Struct. Biol.*, 2016, **72**, 1–10.
61. XChip information, http://www.mitegen.com/product/x-chip/ (accessed on 18.03.2018).
62. A. E. Cohen, S. M. Soltis, A. Gonzalez, L. Aguila, R. Alonso-Mori, C. O. Barnes, E. L. Baxter, W. Brehmer, A. S. Brewster, A. T. Brunger, G. Calero, J. F. Chang, M. Chollet, P. Ehrensberger, T. L. Eriksson, Y. Feng, J. Hattne, B. Hedman, M. Hollenbeck, J. M. Holton, S. Keable, B. K. Kobilka, E. G. Kovaleva, A. C. Kruse, H. T. Lemke, G. Lin, A. Y. Lyubimov, A. Manglik, Mathews II, S. E. McPhillips, S. Nelson, J. W. Peters, N. K. Sauter, C. A. Smith, J. Song, H. P. Stevenson, Y. Tsai, M. Uervirojnangkoorn, V. Vinetsky, S. Wakatsuki, W. I. Weis,

O. A. Zadvornyy, O. B. Zeldin, D. Zhu and K. O. Hodgson, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, 17122–17127.
63. X. Yin, A. Scalia, L. Leroy, C. M. Cuttitta, G. M. Polizzo, D. L. Ericson, C. G. Roessler, O. Campos, M. Y. Ma, R. Agarwal, R. Jackimowicz, M. Allaire, A. M. Orville, R. M. Sweet and A. S. Soares, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2014, **70**, 1177–1189.
64. M. A. Berger, J. H. Decker and Mathews II, *J. Appl. Crystallogr.*, 2010, **43**, 1513–1518.
65. S. Sui, Y. Wang, K. W. Kolewe, V. Srajer, R. Henning, J. D. Schiffman, C. Dimitrakopoulos and S. L. Perry, *Lab Chip*, 2016, **16**, 3082–3096.
66. J. E. Lee, M. L. Fusco and E. O. Saphire, *Nat. Protoc.*, 2009, **4**, 592–604.
67. I. Russo Krauss, A. Merlino, A. Vergara and F. Sica, *Int. J. Mol. Sci.*, 2013, **14**, 11643–11691.
68. S. Guha, S. L. Perry, A. S. Pawate and P. J. Kenis, *Sens. Actuators, B*, 2012, **174**, 1–9.
69. S. L. Perry, S. Guha, A. S. Pawate, R. Henning, I. Kosheleva, V. Srajer, P. J. Kenis and Z. Ren, *J. Appl. Crystallogr.*, 2014, **47**, 1975–1982.
70. D. S. Khvostichenko, J. M. Schieferstein, A. S. Pawate, P. D. Laible and P. J. A. Kenis, *Cryst. Growth Des.*, 2014, **14**, 4886–4890.
71. J. D. Ng, P. J. Clark, R. C. Stevens and P. Kuhn, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2008, **64**, 189–197.
72. CrystalSlide information, http://www.mitegen.com/products/plates/CrystalSlide_UserGuide_E.pdf (accessed on 16.02.2017).
73. CrystalHarp information, https://www.moleculardimensions.com/applications/upload/Crystalharp%20flyer_UK.pdf (accessed on 16.02.2017).
74. B. Zheng, J. D. Tice, L. S. Roach and R. F. Ismagilov, *Angew. Chem., Int. Ed. Engl.*, 2004, **43**, 2508–2511.
75. C. J. Gerdts, M. Elliott, S. Lovell, M. B. Mixon, A. J. Napuli, B. L. Staker, P. Nollert and L. Stewart, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2008, **64**, 1116–1122.
76. J. Christensen, C. J. Gerdts, M. C. Clifton and L. Stewart, *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.*, 2011, **67**, 1022–1026.
77. M. K. Yadav, C. J. Gerdts, R. Sanishvili, W. W. Smith, L. S. Roach, R. F. Ismagilov, P. Kuhn and R. C. Stevens, *J. Appl. Crystallogr.*, 2005, **38**, 900–905.
78. S. V. Akella, A. Mowitz, M. Heymann and S. Fraden, *Cryst. Growth Des.*, 2014, **14**, 4487–4509.
79. D. Axford, P. Aller, J. Sanchez-Weatherby and J. Sandy, *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.*, 2016, **72**, 313–319.
80. R. Schubert, S. Kapis, Y. Gicquel, G. Bourenkov, T. R. Schneider, M. Heymann, C. Betzel and M. Perbandt, *IUCrJ*, 2016, **3**, 393–401.
81. MiTeGen webpage, http://www.mitegen.com/lcp/imisx/ (accessed on 17.02.2017).
82. A. Savill, personal communication.
83. V. Cherezov and M. Caffrey, *Faraday Discuss.*, 2007, **136**, 195.

84. K. Hirata, Towards Automatic Data Collection Pipeline for Membrane Protein Structure Analyses at Beamline BL32XU, Synchrotron Radiation and Instrumentation 2015 conference, 2015.
85. C. G. Roessler, R. Agarwal, M. Allaire, R. Alonso-Mori, B. Andi, J. F. R. Bachega, M. Bommer, A. S. Brewster, M. C. Browne, R. Chatterjee, E. Cho, A. E. Cohen, M. Cowan, S. Datwani, V. L. Davidson, J. Defever, B. Eaton, R. Ellson, Y. Feng, L. P. Ghislain, J. M. Glownia, G. Han, J. Hattne, J. Hellmich, A. Héroux, M. Ibrahim, J. Kern, A. Kuczewski, H. T. Lemke, P. Liu, L. Majlof, W. M. McClintock, S. Myers, S. Nelsen, J. Olechno, A. M. Orville, N. K. Sauter, A. S. Soares, S. M. Soltis, H. Song, R. G. Stearns, R. Tran, Y. Tsai, M. Uervirojnangkoorn, C. M. Wilmot, V. Yachandra, J. Yano, E. T. Yukl, D. Zhu and A. Zouni, *Structure*, 2016, 631–640.
86. C. G. Roessler, A. Kuczewski, R. Stearns, R. Ellson, J. Olechno, A. M. Orville, M. Allaire, A. S. Soares and A. Héroux, *J. Synchrotron Radiat.*, 2013, **20**, 805–808.
87. F. D. Fuller, S. Gul, R. Chatterjee, E. S. Burgie, I. D. Young, H. Lebrette, V. Srinivas, A. S. Brewster, T. Michels-Clark, J. A. Clinger, B. Andi, M. Ibrahim, E. Pastor, C. de Lichtenberg, R. Hussein, C. J. Pollock, M. Zhang, C. A. Stan, T. Kroll, T. Fransson, C. Weninger, M. Kubin, P. Aller, L. Lassalle, P. Brauer, M. D. Miller, M. Amin, S. Koroidov, C. G. Roessler, M. Allaire, R. G. Sierra, P. T. Docker, J. M. Glownia, S. Nelson, J. E. Koglin, D. Zhu, M. Chollet, S. Song, H. Lemke, M. Liang, D. Sokaras, R. Alonso-Mori, A. Zouni, J. Messinger, U. Bergmann, A. K. Boal, J. M. Bollinger Jr., C. Krebs, M. Hogbom, G. N. Phillips Jr., R. D. Vierstra, N. K. Sauter, A. M. Orville, J. Kern, V. K. Yachandra and J. Yano, *Nat. Methods*, 2017, **14**, 443–449.
88. S. Tsujino and T. Tomizaki, *Sci. Rep.*, 2016, **6**, 25558.
89. H. P. Stevenson, D. P. DePonte, A. M. Makhov, J. F. Conway, O. B. Zeldin, S. Boutet, G. Calero and A. E. Cohen, *Philos. Trans. R. Soc. London*, 2014, **369**, 20130322.
90. F. Stellato, D. Oberthür, M. Liang, R. Bean, C. Gati, O. Yefanov, A. Barty, A. Burkhardt, P. Fischer, L. Galli, R. A. Kirian, J. Meyer, S. Panneerselvam, C. H. Yoon, F. Chervinskii, E. Speller, T. A. White, C. Betzel, A. Meents and H. N. Chapman, *IUCrJ*, 2014, **1**, 204–212.
91. A. Y. Lyubimov, T. D. Murray, A. Koehl, *I. E.* Araci, M. Uervirojnangkoorn, O. B. Zeldin, A. E. Cohen, S. M. Soltis, E. L. Baxter, A. S. Brewster, N. K. Sauter, A. T. Brunger and J. M. Berger, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2015, **71**, 928–940.
92. R. G. Sierra, C. Gati, H. Laksmono, E. H. Dao, S. Gul, F. Fuller, J. Kern, R. Chatterjee, M. Ibrahim, A. S. Brewster, I. D. Young, T. Michels-Clark, A. Aquila, M. Liang, M. S. Hunter, J. E. Koglin, S. Boutet, E. A. Junco, B. Hayes, M. J. Bogan, C. Y. Hampton, E. V. Puglisi, N. K. Sauter, C. A. Stan, A. Zouni, J. Yano, V. K. Yachandra, S. M. Soltis, J. D. Puglisi and H. DeMirci, *Nat. Methods*, 2015, **13**, 59–62.

93. R. G. Sierra, H. Laksmono, J. Kern, R. Tran, J. Hattne, R. Alonso-Mori, B. Lassalle-Kaiser, C. Glöckner, J. Hellmich, D. W. Schafer, N. Echols, R. J. Gildea, R. W. Grosse-Kunstleve, J. Sellberg, T. A. McQueen, A. R. Fry, M. M. Messerschmidt, A. Miahnahri, M. M. Seibert, C. Y. Hampton, D. Starodub, N. D. Loh, D. Sokaras, T.-C. Weng, P. H. Zwart, P. Glatzel, D. Milathianaki, W. E. White, P. D. Adams, G. J. Williams, S. Boutet, A. Zouni, J. Messinger, N. K. Sauter, U. Bergmann, J. Yano, V. K. Yachandra and M. J. Bogan, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2012, **68**, 1584–1587.
94. P. Nogly, V. Panneels, G. Nelson, C. Gati, T. Kimura, C. Milne, D. Milathianaki, M. Kubo, W. Wu, C. Conrad, J. Coe, R. Bean, Y. Zhao, P. Båth, R. Dods, R. Harimoorthy, K. R. Beyerlein, J. Rheinberger, D. James, D. DePonte, C. Li, L. Sala, G. J. Williams, M. S. Hunter, J. E. Koglin, P. Berntsen, E. Nango, S. Iwata, H. N. Chapman, P. Fromme, M. Frank, R. Abela, S. Boutet, A. Barty, T. A. White, U. Weierstall, J. Spence, R. Neutze, G. Schertler and J. Standfuss, *Nat. Commun.*, 2016, **7**, 12314.
95. S. Botha, K. Nass, T. R. Barends, W. Kabsch, B. Latz, F. Dworkowski, L. Foucar, E. Panepucci, M. Wang, R. L. Shoeman, I. Schlichting and R. B. Doak, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2015, **71**, 387–397.
96. U. Weierstall, *Philos. Trans. R. Soc. London*, 2014, **369**, 20130337.
97. M. Boudes, D. Garriga, A. Fryga, T. Caradoc-Davies and F. Coulibaly, *Acta Crystallogr., D: Struct. Biol.*, 2016, **72**, 576–585.
98. K. R. Beyerlein, D. Dierksmeyer, V. Mariani, M. Kuhn, I. Sarrou, A. Ottaviano, S. Awel, J. Knoska, S. Fuglerud, O. Jönsson, S. Stern, M. O. Wiedorn, O. Yefanov, L. Adriano, R. Bean, A. Burkhardt, P. Fischer, M. Heymann, D. A. Horke, K. E. J. Jungnickel, E. Kovaleva, O. Lorbeer, M. Metz, J. Meyer, A. Morgan, K. Pande, S. Panneerselvam, C. Seuring, A. Tolstikova, J. Lieske, S. Aplin, M. Roessle, T. A. White, H. N. Chapman, A. Meents and D. Oberthuer, *IUCrJ*, 2017, **4**, 769–777.

CHAPTER 2

Delivery of GPCR Crystals for Serial Femtosecond Crystallography

E. E. ABOLA,[a] U. WEIERSTALL,[b] W. LIU[c] AND V. CHEREZOV*[a]

[a] University of Southern California, Bridge Institute, Department of Chemistry, 1002 Childs Way, Los Angeles 90089, USA; [b] Arizona State University, Department of Physics, Tempe, AZ 85287-1504, USA; [c] Biodesign Center for Applied Structural Discovery, Biodesign Institute, School of Molecular Sciences, Arizona State University, Tempe, AZ 85287-1504, USA
*Email: cherezov@usc.edu

2.1 Introduction

G protein-coupled receptors (GPCRs) comprise the largest protein superfamily in humans, containing about 800 members which are classified into 5 major families: rhodopsin, secretin, glutamate, frizzled, and adhesion, based on sequence similarity within their 7-transmembrane (7TM) domain. GPCRs reside in the plasma membrane and transmit signals inside the cells in response to a variety of extracellular stimuli, ranging from photons, ions and small molecules to peptides and proteins. Depending on their chemical structures, different ligands have variable signaling efficacies (*i.e.* acting as agonists, biased agonists, antagonists, inverse agonists, or allosteric modulators). Signal transmission occurs primarily through coupling to heterotrimeric G proteins which, in turn, activate downstream effectors triggering cascades of cellular and physiological responses. GPCR-mediated

Chemical Biology No. 8
Protein Crystallography: Challenges and Practical Solutions
Edited by Konstantinos Beis and Gwyndaf Evans
© The Royal Society of Chemistry 2018
Published by the Royal Society of Chemistry, www.rsc.org

signaling pathways play a key role in all physiological systems as well as pathophysiological conditions including cancer, immune disorders, cardiovascular diseases, metabolic disorders (*e.g.* obesity, diabetes), pain and addiction, *etc.* Such a central role in crucial physiological processes makes members of the GPCR superfamily attractive drug targets; over 30% of all prescription drugs on the market act *via* these receptors.[1,2]

Despite a number of recent breakthroughs in GPCR expression, stabilization, and crystallography, their structure determination remains challenging primarily because they adopt multiple conformations as required by their biological function. They readily switch between inactive and active states depending on bound ligands and effector proteins as well as their inherent basal activity levels. Currently, over 100 structures for more than 40 distinct GPCRs have been solved with crystallization done mostly in lipidic cubic phase (LCP). LCP serves as a native-like membrane mimetic supporting crystal nucleation and growth. LCP crystallization was introduced in 1996 by E. Landau and J. Rosenbusch,[3] matured over the last 20 years,[4] and is currently becoming increasingly popular among structural biologists working with membrane proteins. One of the limitations toward obtaining high-resolution GPCR structures is the relatively small size of LCP-grown crystals. Membrane protein crystals grown in LCP have type I packing with both hydrophilic and hydrophobic crystal contacts and a low solvent content. Such crystals typically have a very high order capable of diffracting to a high resolution but often grow to limited size (~several microns). The amount of crystallographic information that can be obtained from well-ordered microcrystals using state-of-the-art microfocus beamlines at synchrotron sources is also strongly limited by radiation damage. Current practices include collecting small wedges of data of few degrees from many frozen crystals and merging them together in a dataset.[5] This procedure requires extensive optimization of crystal growth, harvesting hundreds of crystals, aligning them by beam rastering, and using special algorithms for data processing, all of which involve substantial commitments in time and effort, while at the end still having to tolerate a certain amount of radiation damage.

All of these time-consuming steps and, most importantly, the effects of radiation damage can be essentially avoided by taking advantage of new generation X-ray sources based on X-ray free electron lasers (XFELs). XFELs produce high brilliance X-ray pulses of a few femtoseconds long, allowing for high-resolution data collection on microcrystals before damage occurs. The first proofs of principle for Serial Femtosecond Crystallography (SFX) have been demonstrated by collecting and processing data at XFELs using continuous streams of microcrystals of soluble proteins (lysozyme, cathepsin B),[6,7] and of membrane proteins grown either in detergent solution (photosystem I)[8] or in a lipidic sponge phase (reaction complex).[9] A serious drawback of delivering microcrystals into an XFEL beam using a liquid-jet device employed in these studies is the relatively high flow-rate ($10\ \mu L\ min^{-1}$), which, coupled with the 120 Hz pulse repetition rate, requires

tens to hundreds of milligrams of crystallized protein to collect a full dataset. Alternatively, the high viscosity and gel-like consistency of LCP allows the reduction and control of the flow rate over a wide range, which ensures more efficient use of microcrystals, and decreases the required amount of crystallized protein material by about 2–3 orders of magnitude. SFX is therefore especially suited for data collection on highly ordered microcrystals of membrane proteins grown in LCP and is revolutionizing the field by accelerating the pace of membrane protein structural determination. Large-scale production of GPCR crystals in LCP and their delivery to an XFEL beam using an extruded stream of LCP are discussed in this chapter.

Structural biology of GPCRs is currently just in its initial phase; over 40 unique receptor structures are available, which amounts to about 10% of all known human non-olfactory receptors. Questions about subtype specificity as well as polypharmacology of receptors are among the leading issues that have to be addressed. Although some insights into the activation process have been achieved through a limited number of structural studies, important questions remain, most notably the molecular basis for the selection of a specific G protein. The recent structure determination of a major signaling complex between rhodopsin and arrestin[10] demonstrated the potential power of SFX and establishes a framework that can be used to determine other challenging structures, to validate the results, as well as to answer important biological questions.

2.2 Process Overview

Studies leading to the determination of the structures of the β_2-adrenergic receptor[11,12] and of the adenosine A_{2A} receptor,[13] the first two structures of GPCRs bound to a diffusible ligand, generated a blueprint for a process, which has now been optimized to operate in high-thoughput, working in parallel with several constructs and even several targets. Within this process, attempts are made to achieve several intermediate goals: the large-scale production of stable constructs and their characterization, the crystallization of these constructs, the acquisition of high quality diffraction data, and finally the determination and refinement of the structure. These goals are commonly achieved through an iterative process using protocols and technologies described in a number of publications,[14–24] including those that report new structures.[11,13,25–42] This pipeline includes steps necessary for generating and delivering samples for SFX experiments. GPCRs are highly dynamic and unstable molecules, which typically do not crystallize in their wild type form and have to be stabilized in a particular conformational state *via* protein engineering approaches. The process requires design, production, and testing of a large number of protein-constructs (typically a few hundred) and screening of many ligands to identify the construct–ligand combination that can be crystallized. The pipeline relies on a number of metrics measured at specific process steps, which have proven to be an important guide in reducing processing work (and cost) on nonproductive

constructs. These criteria include expression levels >0.25 mg L^{-1}, near native binding activity as measured by a radioligand competitive inhibition assay, purity >90% (SDS PAGE), low aggregation (SEC), high receptor unfolding temperature Tm >60 °C,[14] high Tm in LCP (LCP-Tm),[17] high receptor diffusion in LCP (LCP-FRAP),[19] and the ability to generate crystals that diffract to better than 3.5 Å. In cases where ligands are not available, either antibodies (Fab fragments) or nanobodies are used to attempt stabilization of a unique conformation.[43]

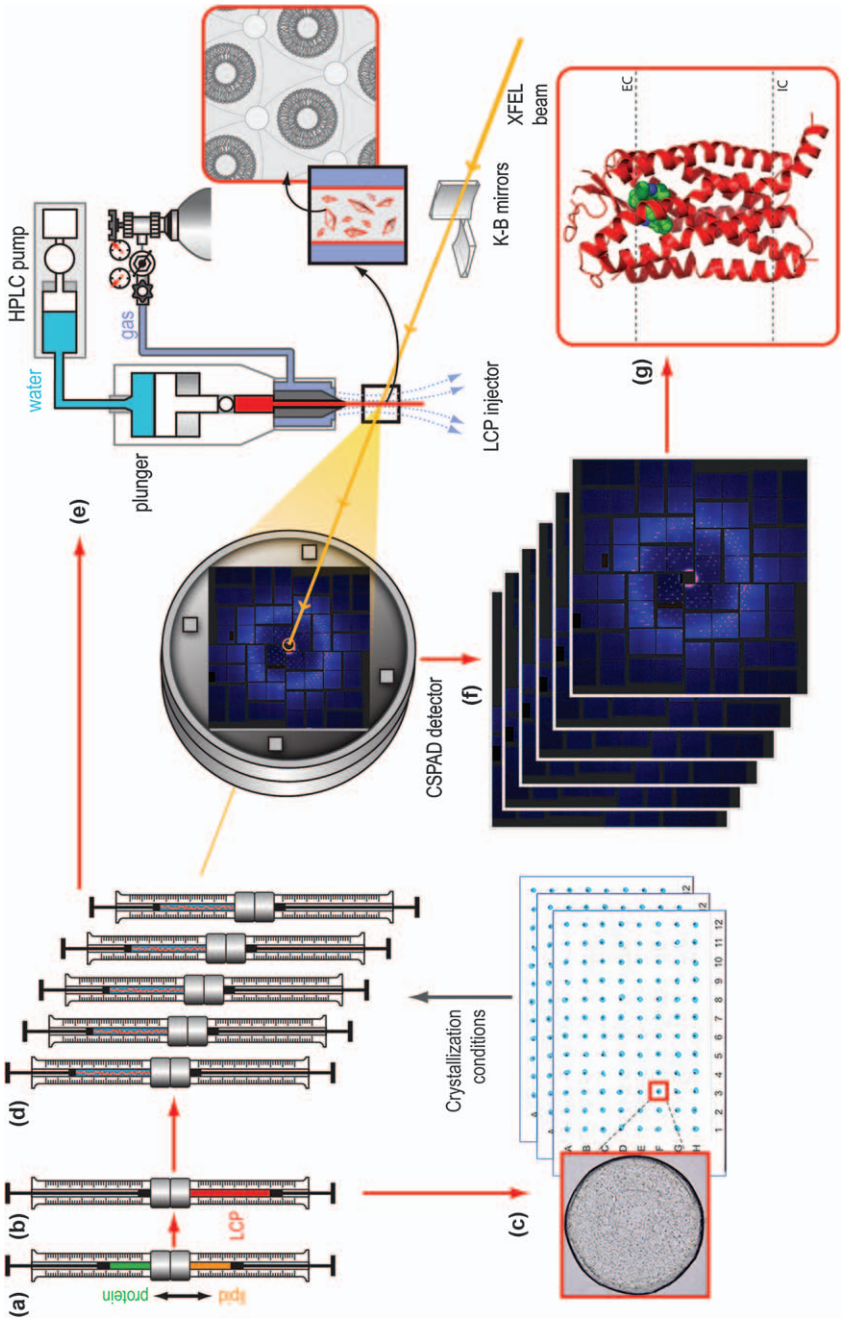
Once microcrystals become available, the workflow for LCP-SFX experiments shown in Figure 2.1 can be executed. No further optimization to produce large crystals is needed. The high brilliance of XFEL sources enables the use of randomly oriented micron-sized crystals (<10 μm), which are embedded in an extruded stream of LCP and are intersected by the pulsed X-ray beam. The concentration of crystals in LCP has to be adjusted so that on average each X-ray pulse diffracts from single crystals. The short duration of the X-ray pulses, <50 fs, allows the diffracted beam to escape the crystal before photoionization damage is established; a phenomenon now termed ''diffraction before destruction''. Finally, specialized data processing software is used to generate structure factors integrated from multiple measurements of partial reflections.[44] Standard approaches, such as molecular replacement,[22] anomalous diffraction[45] and isomorphous replacement,[46] are then used to determine and refine the structure.

2.3 Achieving Major Milestones

In the following sections, we provide further details on the processes used in attempting to achieve these intermediate goals. We attempt to provide examples, discuss critical issues and possible pitfalls, and remedial steps that could be taken to address them.

2.3.1 Large-scale Production of Stable Receptor Constructs

The objective is to produce sufficiently large volumes of GPCR samples that can be used successfully in crystallization and structure determination studies. This is achieved by addressing several underlying issues, namely the need for high expression levels, and the need to stabilize the receptor outside its native membrane environment as well as to overcome its disposition to adopting multiple conformations. Receptor dynamics is essential for carrying out its biological role but is detrimental to growing high quality crystals. Protein engineering approaches have been found to be effective in addressing these issues and have been integrated into an iterative multistep process with feedback loops, which is optimized for each receptor or receptor subfamily with the aim to produce high quality samples for structural biology studies.[21,47,48] The main sample production process steps are (A) design of constructs, (B) overexpression in either insect or mammalian


(a)
protein
lipid
(b)
LCP
(c)
Crystallization conditions
(d)
(e)
HPLC pump
water
gas
plunger
LCP injector
K-B mirrors
XFEL beam
CSPAD detector
(f)
(g)
EC
IC

cells, and (C) purification and characterization. These steps are executed until samples achieving or exceeding our threshold metrics are generated.

2.3.1.1 Construct Design

Protein engineering for structural studies is done to address a number of issues: (A) to improve expression levels, (B) to alter post-translational modification sites such as removing glycosylation sites (sources of inhomogeneity), (C) to stabilize highly dynamic domains, such as the third intracellular loop (ICL3) in Class A GPCRs, by introducing protein fusions that do not significantly alter receptor function while providing a structured polar surface that increases crystallizability, (D) to reduce possible protease cleavage sites, such as the removal of flexible N- and C-termini, and (E) to improve thermostability by introducing site-directed mutations.

The best construct for determining the structure of a receptor is currently arrived at by a trial and error approach. Initial studies of receptors, for which structures are not available, are therefore aimed at developing an early understanding of their behavior, particularly their inherent stability. This normally includes the choice of promoters, placements of affinity tags for purification, N- and C-terminal truncations, and introduction of fusion protein using the GPCR Protein Fusion Partner Toolchest.[20] For receptors with large extracellular domains such as Class B, C, F or adhesion receptors, structure determination studies may opt to initially limit studies to the 7TM domain. Molecular modeling is generally employed to rationally design fusion sites and point mutations for stability and crystallographic studies. In our laboratory, modeling is performed using a comprehensive set of algorithms developed and implemented by the GPCR network, based on an internal coordinate mechanics platform[49] for conformational energy optimization and flexible molecular docking, as well as molecular dynamics simulations in lipidic membrane environment.[50,51]

The first successful and most commonly used fusion protein is T4 lysozyme, which has been used in over half of the solved structures. Another common fusion partner is a thermostabilized apocytochrome b_{562}RIL (BRIL), which has been used for the highest resolution GPCR structures

Figure 2.1 LCP-SFX workflow: (a) protein solution is mixed with lipid using a syringe mixer to (b) reconstitute protein in LCP; (c) high-throughput LCP crystallization trials are conducted in 96-well sandwich plates to identify crystal hits; (d) LCP crystallization is set up in syringes using conditions identified by the high-throughput crystal screening; (e) crystals grown in syringes are consolidated and loaded in the LCP injector and streamed across an XFEL beam for SFX data collection; (f) diffraction images are collected and processed to be used for (g) structure determination and refinement. Adapted with permission from W. Liu, D. Wacker, C. Wang, E. Abola and V. Cherezov, *Philos. Trans. Royal Soc. B*, 2014, **369**, 20130314; DOI: 10.1098/rstb.2013.0314 © 2014 The Author(s) Published by the Royal Society. All rights reserved.

determined at 1.7 and 1.8 Å.[38,52,53] Other successfully used fusion proteins include rubredoxin[35,39] and catalytic domain of PGS (*Pyrococcus abyssi* glycogen synthase).[54–56] The majority of crystallized constructs have the protein fusion inserted into the ICL3 or, in case of several non-class A GPCRs, into ICL2,[57–59] while others have it attached to the N-terminus. When possible, attaching the fusion protein to the N-terminus, as it was done, for example, for the nociceptin/orphanin FQ receptor (NOP)[29] or the angiotensin II receptor type 1,[41] has the advantage of being able to use signaling assays to test for functional activity. Typically a panel of several small, stable proteins are tested as N-terminal and ICL3 fusion partner candidates[20] to decide which fusion constructs to optimize based on the metrics discussed below.

2.3.1.2 Overexpression in a Eukaryotic System

The ultimate goal of overexpression is the large-scale production of receptor constructs that can be used for crystallization studies. This requires the identification of the best expressing constructs as well as the best conditions and approaches to use. The immediate goal is to be able to meet or exceed the threshold metric of producing >0.25 mg of protein per liter of biomass. Ultimately, downstream processing steps will be used to identify constructs for production at a scale needed to carry out SFX studies. Several expression systems have now been used to produce GPCR samples for crystallization. All but five of them used insect cells, while the others used *E. coli*,[60] yeast,[27] and HEK293 mammalian cells.[10,61,62] Here we provide details of our protocols using the baculovirus expression system.

The cDNA for each construct is synthesized, usually through commercial providers. Sequence-verified constructs are then received in a shuttle vector, which is amplified and digested with flanking restriction sites, and ligated into four previously constructed expression cassettes cloned into the pFastBacI transfer vector, which is commercially available from Invitrogen.

We have established four vectors as the backbone of our expression screening efforts using the Bac-to-Bac system (Invitrogen). Vectors 1 and 2 are under the control of the standard late polyhedrin promoter whereas vectors 3 and 4 are under control of the early to late GP64 promoter. The GP64 promoter was cloned into pFastBac1 (Invitrogen) from pBAC5 (Novagen) to explore the effects of a weaker promoter on overall protein expression levels and resulting protein quality. Vectors 1 and 3 have an N-terminal FLAG (Sigma) epitope tag whereas vectors 2 and 4 contain a C-terminal FLAG tag. The expression cassettes utilized here were designed to maximize the probability of success with any given GPCR target while easing downstream analysis of expression and stability through the incorporation of purification and detection tags. This vector set has generated high success rates, showing high levels of cell surface expression for the membrane proteins that we have evaluated. The gene of interest is subcloned into the cassette immediately downstream by virtue of the 5′ AscI and

3′ FseI restriction sites that are followed by a PreScission protease cleavage sequence, a 10xHis tag, and an optional FLAG epitope tag. In essence, the difference between the two cassettes is in the placement of the FLAG epitope tag either upstream of the gene of interest (vectors 1 and 3) or downstream (vectors 2 and 4). Receptor overexpression is then carried out using a baculovirus (BV) expression system in *Spodoptera frugiperda* (*Sf9*) insect cells.

In general, we carry out expression studies at three scales: (A) small scale (~4 mL), to screen expression levels and to survey conditions that may lead to higher expression of more stable constructs, such as conducting preliminary assessment of the effect of additives (*e.g.* ligands, cholesterol) on the expression levels, (B) medium scale (40–250 mL), to fully characterize expressed constructs, and (C) large scale (≥1 L), to produce constructs for crystallization and data collection. Constructs that do not meet our threshold metrics are generally not selected for further studies. This approach enables parallel studies on multiple targets and/or constructs. In addition, decisions on proceeding with a construct can be made as early as possible in the process, using the least amount of resources (*i.e.* manpower and reagents). In most cases that we have worked on, conditions/parameters used in small-scale studies have proven to be applicable to larger scale studies thus obviating the need for additional extensive parameter optimization studies.[63]

Characterization of Expression: Initial characterization of expression levels is performed with 100 μL of the expression cell culture using flow cytometry analysis which detects a common epitope expressed on the N-terminus of the receptors such as the FLAG tag, using primary antibodies coupled to an appropriate Zenon Fab fragment (Invitrogen). The mean fluorescence intensity (MFI) weighted by expressing cell population is used directly on intact cells to estimate cell surface protein expression.[63] In addition, total protein expression is estimated either by probing permeabilized cells with the fluorescently labeled antibody or by utilizing a C-terminally fused fluorescent tag such as GFP when available. The ratio of cell surface expression to total expression (MFI_{np}/MFI_{p}) is used as an approximation of protein integrity,[64] also providing insight on production issues such as protein trafficking. Depending on the expression cassette used, there are instances where the epitope tag is not located on the N-terminus of the protein and total protein (MFI_{p}) is then used as an approximation of expression levels to select constructs for further processing and scale up to medium scale expression. Each expression run is assayed for desired protein product by small-scale extraction, immobilized metal affinity chromatography (IMAC)-based purification, and SDS-PAGE analysis with western blotting.

Expression Optimization: When constructs are identified that express at medium to high levels, several optimization steps are carried out to maximize expression levels and reduce the overall cost of production. There are a number of reported incidences in the literature where addition of a ligand to the expression medium has enhanced the levels of properly folded protein, essentially by further stabilizing the folded conformation of the receptor.[65,66]

Similarly, it has long been known that many of the family A GPCRs tend to be affected rather dramatically by the presence of cholesterol in the plasma membrane. Interestingly, there is a striking lack of sterol compounds in the plasma membrane of *Sf9* cells raising the question of what effect incorporation of cholesterol may have on the expression levels of receptors that are known to interact favorably with sterols. We have observed significant increase in expression levels for many different receptors upon incorporation of cholesterol prior to viral infection, a finding that is not surprising given the effect of cholesterol on the binding properties of certain receptors as well as the recently appreciated specific cholesterol binding site in 20–40% of human class A GPCRs.[67–71] In addition to these two classes of chaperones, the traditional issues of multiplicity of infection (MOI), time of harvest post infection, and host cell strain are adjusted to maximize expression levels.

2.3.1.3 Purification and Stabilization

The goal is to arrive at a sample purification protocol that leads to the large-scale production of receptor constructs for crystallization studies. The relevant metrics that have to be met include purity >90% (SDS PAGE), low aggregation (SEC), high receptor unfolding temperature Tm > 60 °C,[14] high Tm in LCP (LCP-Tm),[17] and high diffusion in LCP (LCP-FRAP).[19,72] Membrane extraction and solubilization in detergent/cholesteryl hemisuccinate (CHS) micelles is followed by IMAC-based purification that has proven to be an adequate starting point for most receptors studied to date. However, the protocol must be optimized extensively for each construct to maximize purity and stability of the final sample.[11,13,16,25–40,73] At high expression levels of recoverable protein, purification can generally be achieved with a single affinity column chromatography step utilizing the 10xHis tag with an IMAC protocol.

In addition to assays for purity, monodispersity, and activity, protein constructs are analyzed for additional properties that have been found to be important for crystallization. The stability of the protein both in detergent micelles[14] and in LCP[17] is measured and optimized to ensure high quality protein is being delivered to crystallization trials, and that the lipid environment of LCP is conducive to the stability of the receptor. In addition, as a final pre-crystallization screen, constructs are tested for their ability to diffuse in the lipid matrix used for crystallization trials.[19]

Selection of Detergent: Screening for the best detergent and additives that lead to the highest yield of pure, stable, and functional receptor is needed to identify those that can efficiently extract the protein from the membrane yet also be mild enough to maintain protein stability. If necessary, one detergent may be selected for membrane extraction and then be exchanged with another one for crystallization. Tests are also carried out to assess the effects of high detergent concentration on the stability of LCP.

In aqueous solutions detergents spontaneously assemble into micelles with hydrophobic tails hidden in the micelle core and polar heads exposed

to the solution. This process is described by an important physicochemical parameter, known as the critical micelle concentration (CMC), which corresponds to the minimal detergent concentration required for the formation of micelles. At concentrations higher than the CMC, the detergents exist in a monomer–micelle equilibrium, and the concentration of free monomer detergent molecules in solution remains essentially constant.[74] As a rule-of-a-thumb, detergents are used at concentrations 10–20×CMC for extraction and at least 1.5×CMC for protein purification. While CMC is mostly defined by the chemical structure of the detergent molecule, it also depends on environmental factors, such as temperature and ionic strength of the buffer. Detergent micelles comprise a few dozen to several hundred detergent molecules, and the number of molecules in a micelle (aggregation number) constitutes another important detergent parameter.[74]

For GPCRs, the mixed micelle system of *n*-dodecyl-β-D-maltoside (DDM) and CHS has thus far been the most frequently used combination[75] to extract GPCRs from the membrane and purify them, with final concentrations ranging from 0.5% to 1% and 0.03% to 0.2% (w/v), respectively. Other common detergents for GPCR purification are maltose-neopentyl glycols,[76] such as lauryl maltose-neopentyl glycol (LMNG). Detergent solutions for purification have often been formulated with 50–1000 mM sodium chloride and buffered at pH 6.0–8.0 with an appropriate buffering system, commonly MES, Bis-Tris Propane, HEPES or Tris-HCl.

Selection of LCP-host lipid: Once stable purified receptor constructs are available, a search is initiated to find the best lipid for LCP (*in meso*) crystallization. Most of the lipids suitable for LCP crystallization belong to a large class of monounsaturated monoacylglycerols (*N.T* MAGs, where *N* is the number of hydrocarbons between the ester bond and the double bond, and *T* is the number of hydrocarbons between the double bond and the end of the hydrophobic chain). Monoolein (9.9 MAG) has been by far the most successful lipid used for LCP crystallization to date;[77] however, several other MAGs, such as 7.7 MAG, 7.9 MAG and 9.7 MAG, have proven to be superior to monoolein in certain cases. LCP samples made of different host lipids have different bilayer thicknesses and water channel diameters, and they respond differently to precipitant solutions, which may affect the stability, diffusion and interaction of the reconstituted membrane proteins. In some cases, doping LCP with phospholipids, cholesterol, or other native lipids can improve crystal growth and quality. Most of the GPCRs have been crystallized in LCP made of a mixture of monoolein with 10% (wt/wt) cholesterol.

Selected protein constructs are assessed after reconstitution in LCP for properties that have been found to be important for crystallization, including stability of the protein in LCP and the impact of lipid environment.[17] LCP crystallization mechanisms require that the reconstituted protein undergoes relatively fast diffusion inside the lipidic matrix; construct–ligand combinations are therefore tested for their ability to diffuse in a large number of precipitant conditions using high-throughput LCP-FRAP assay.[19]

2.3.2 Crystallization of Receptor Constructs

Crystallization in LCP often results in showers of well-ordered but small micron-sized crystals, which are ideally suited for structure determination by SFX. Such micron-sized crystals are delivered to an XFEL beam inside of the same LCP matrix in which they were grown using special protocols.[24] Using a highly purified and stable receptor construct, the goal is to generate at least 50 μL of LCP filled uniformly and at high density with microcrystals, each less than 10 μm in size. A two-stage process is used to achieve this goal (Figure 2.1). The first stage uses high-throughput LCP nanovolume crystallization screening to identify conditions that lead to a high rate of nucleation and to the formation of a large number of microcrystals. This step follows protocols that are currently used for *in meso* crystallization of membrane proteins[15,18,78,79] with robotic systems. The second stage uses a multistep process that we developed for scaling-up crystallization volumes by 1000 times.[24,80] This scale-up process addresses issues that arise due to the gel-like texture of LCP and the slow diffusion of soluble components within it, which could result in transient concentration gradients, the extent and time dependence of which strongly depend on the volume and shape of the LCP crystallization bolus.

The process begins with reconstitution of the membrane protein of interest in LCP, and then it proceeds to setting up the crystallization experiments in gas-tight syringes (Hamilton). Protein-laden LCP is extruded from one syringe into a precipitant solution in several other syringes as a ~400 μm (diameter) string, closely mimicking the geometry used during initial high-throughput crystallization experiments in glass sandwich plates, but in a much larger LCP volume. The syringes are then sealed and incubated at 20 °C to allow the crystals to nucleate and grow. At this stage, the samples can be transported to the beamline inside a Greenbox thermal management system (ThermoSafe Brands) equilibrated at 20 °C. After microcrystals have appeared and reached the desired size, the precipitant solution is discarded and the LCP sample with embedded microcrystals is consolidated into one syringe. The lipid composition of the sample is then adjusted, if necessary, to account for lipid-phase behavior that varies with specific lipid used.[81–83] The initial reconstitution of a protein into the lipid bilayer of LCP is performed at room temperature and at close to the maximal hydration capacity of LCP, which is just below the full hydration boundary. For most MAGs, this corresponds to the aqueous content of 40–50% (wt/wt). When a membrane protein sample that has been solubilized in detergent is reconstituted in LCP and then overlaid with a precipitant solution, the phase boundaries can shift, and in some cases LCP can transform into a different mesophase.[84,85] It is therefore important to verify that the final mesophase, in which microcrystals are grown, is an LCP, which is indicated by a gel-like appearance and the lack of birefringence, except for the birefringence that comes from microcrystals. Finally, the crystal density and size distribution are characterized and optimized, and the sample is transferred into an LCP

injector for SFX data collection. To minimize the chance of clogging the injector nozzle, the lipid, protein, and precipitant solutions should be prefiltered, syringes and needles should be thoroughly cleaned, and all manipulations in which the sample is exposed to air should preferably be performed inside a clean-room flow hood.

When LCP containing microcrystals is extruded into vacuum, rapid evaporative cooling can induce the transition into a crystalline Lc phase. This is a common occurrence when monoolein is used as a host lipid because the equilibrium LCP-to-Lc transition temperature for monoolein is 18 °C,[83] which is only a few degrees below room temperature. The transition to Lc phase may physically damage the crystals, and, moreover, results in excessive powder diffraction from the crystalline lipid that can oversaturate the detector. We have established that this problem can be eliminated by using either 7.9 MAG or 9.7 MAG as a host lipid.[23] When crystal growth requires LCP made of 9.9 MAG, the problem of phase change upon extrusion in vacuum can be overcome by adding 7.9 MAG after crystal growth, as described in published protocols.[24] When LCP extrusion is performed at ambient pressure the use of 7.9 MAG is not required.

2.3.3 SFX Data Collection

2.3.3.1 Experimental Layout

All SFX data of GPCR structures published to date have been collected at the Coherent X-ray Imaging (CXI)[86,87] beamline at the Linac Coherent Light Source (LCLS). The LCLS is part of the SLAC National Accelerator Laboratory in Menlo Park, California, operated by Stanford University on behalf of the Department of Energy. The CXI beamline is equipped with a Cornel-SLAC Pixel Array Detector (CSPAD)[88] capable of data acquisition at the LCLS pulse repetition rate of 120 Hz, a vacuum sample chamber, and a pair of Kirkpatrick–Baez mirrors focusing the XFEL beam to a diameter of ~1 μm on the sample.

2.3.3.2 LCP Injector and Crystal Delivery

Due to the XFEL pulse repetition rate of 120 Hz (at the LCLS, 60 Hz at SACLA), and since each X-ray pulse completely destroys the sample, the GPCR microcrystals have to be quickly replenished. Each X-ray pulse should illuminate on average only one crystal and background scattering should be kept low. These requirements can be fullfilled by extruding the LCP in a thin stream of 20–50 micron in diameter across the X-ray beam. The LCP extrusion is done either in vacuum for reduced background scattering or in a helium atmosphere. A special device has been developed[23] that allows the extrusion of such a high viscosity gel-like material as a thin stream into vacuum or ambient. The LCP injector (Figure 2.2a) consists of a sample reservoir and a hydraulic stage, which is driven by an HPLC pump. Water

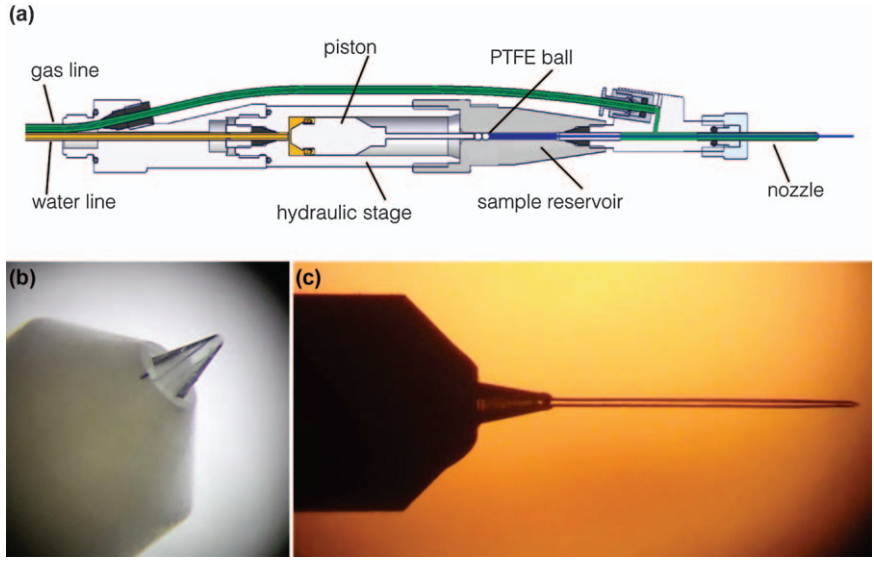


Figure 2.2 LCP extrusion injector. (a) Cross section view of the LCP injector. The piston is actuated by pressurized water at constant flow rate from an HPLC pump (yellow line). The hydraulic stage amplifies the pressure from the HPLC pump and the piston moves two PTFE balls in the LCP sample (blue) reservoir. The balls act as plunger and push the sample into a fused silica capillary with an inner diameter of 50 μm. The sample exits the capillary as a continuous stream into vacuum or air and is kept on axis by co-flowing helium or nitrogen gas (green line). Reprinted from ref. 23. Copyright 2014 Macmillan Publishers Limited. All rights reserved. (b) The coned end of the fused silica capillary (50 μm ID) protrudes out of the square gas aperture made of ceramic. In operation, pressurized gas exits through the corners of the square aperture surrounding the glass capillary cone. (c) LCP extrusion with the LCP injector. The flow rate can be adjusted within the range of 20–2000 nL min^{-1}, which corresponds to LCP flow speeds of 0.17 mm s^{-1} to 17.0 mm s^{-1} for a stream with 50 μm diameter.

from an HPLC pump actuates a piston, which, due to its different diameter at both ends, amplifies the HPLC pump pressure by a factor of 34. The smaller diameter side of the piston drives a PTFE ball, which pushes the LCP and extrudes it through a fused silica capillary with 20–50 μm inner diameter. The PTFE ball is inserted into the sample reservoir bore, which has a slightly smaller diameter than the ball. The ball gets compressed upon insertion into the reservoir and seals against the high pressure needed to drive the LCP through the narrow capillary bore. The distal end of the glass capillary is ground to a cone and protrudes out of a square aperture. This aperture tube is made out of ceramic by injection molding (Figure 2.2b). The LCP is extruded from the coned end of the capillary into vacuum or helium atmosphere (Figure 2.2c), and it would normally curl up on extrusion similar to toothpaste pressed out of a tube. To prevent curling, pressurized helium

gas is exiting through the square aperture and co-flowing with the LCP stream. Shear forces from the co-flowing gas keep the stream on axis and centered onto the intersection with the X-ray beam. The pressure necessary for extrusion of LCP through a 50 μm inner diameter capillary is typically 2000–3000 psi. The X-ray pulses hit the free stream about 100–200 μm downstream of the nozzle end, at which point the pressure on the sample is already relaxed back to ambient conditions.

The reservoir of the LCP injector is loaded with crystal-laden LCP directly from the syringe in which the crystals where grown. Reservoirs with volumes of 25 μL, 40 μL, and 120 μL are available. The LCP flow rate can be adjusted at the HPLC pump to ensure that there is enough sample replenished between XFEL pulses to avoid exposing damaged material to the next pulse. The necessary distance between X-ray hits on the LCP stream depends on the X-ray beam diameter and pulse energy, which typically results (for a 1 μm beam at the LCLS) in a distance of between 10 and 30 μm, *i.e.* a flow speed of about 1.2–3.6 mm sec^{-1}. Flow rates of 20–2000 nL min^{-1} can be achieved by adjusting the flow rate at the HPLC pump, keeping in mind that the flow rate at the sample equals the pump flow rate divided by the pressure amplification factor of the injector (34 for 25 uL reservoirs and 14 for 40 and 120 uL reservoirs). For typical GPCR structure determination, the flow rate at the LCLS is set to about 200 nL min^{-1},[22] whereas for time-resolved pump-probe measurements the flow rate was increased to 2 μL min^{-1},[89] to avoid double exposures and to record three dark exposures in between the laser pumped exposures.

2.3.3.3 SFX Data Processing

Due to the fast XFEL pulse repetition rate (120 Hz at LCLS), a typical LCP-SFX experiment quickly generates hundreds of thousands to millions of detector images, many of which are blank since not every XFEL pulse hits a crystal. The first step in data processing therefore is to identify useful patterns that contain Bragg reflections, referred to as hits. The program Cheetah[90] evaluates each detector image on metrics such as the number of Bragg peaks, maximum resolution, and detector oversaturation levels, while simultaneously performing pre-processing steps necessary for downstream analysis. Additionally, some preliminary data quality metrics can be obtained through the generation of virtual powder diffraction patterns and statistics on hit rate and resolution.

Once hits have been identified, the data are further processed using one of the available programs specifically developed or adapted for SFX, such as CrystFEL,[44] cctbx.xfel,[91] or nXDS.[92] The challenges of SFX data processing include the inherent partiality of all recorded reflections (no rotation), fluctuations in spectrum and intensity between XFEL pulses produced by self-amplified spontaneous emmision (SASE), variability in size, mosaicity, lattice parameters, and diffraction quality of different crystals, as well as the indexing ambiguity affecting certain space groups.[93,94] Partial reflections are

typically merged using a Monte Carlo approach,[95] which relies on high redundancy of individual observations. Various enhancements improving accuracy of obtained structure factors and reducing the number of required indexed images per dataset emerged in recent years, including refinement of the detector geometry,[96] inclusion of linear and Debye–Waller scaling terms,[97] refinement of crystal parameters before integration,[91,98] and estimation of reflection partialities and post-refinement.[99,100]

2.4 Summary of Successful GPCR Structural Studies at XFELs

The LCP-SFX approach has proven highly successful, with structures of seven different GPCRs solved within the last 3.5 years (Figure 2.3 and Table 2.1). The method was first developed and validated in 2013 using the human serotonin 5-HT_{2B} receptor in complex with ergotamine (ERG) as a test protein.[22] The 2.8 Å resolution XFEL structure of 5-HT_{2B}/ERG determined at

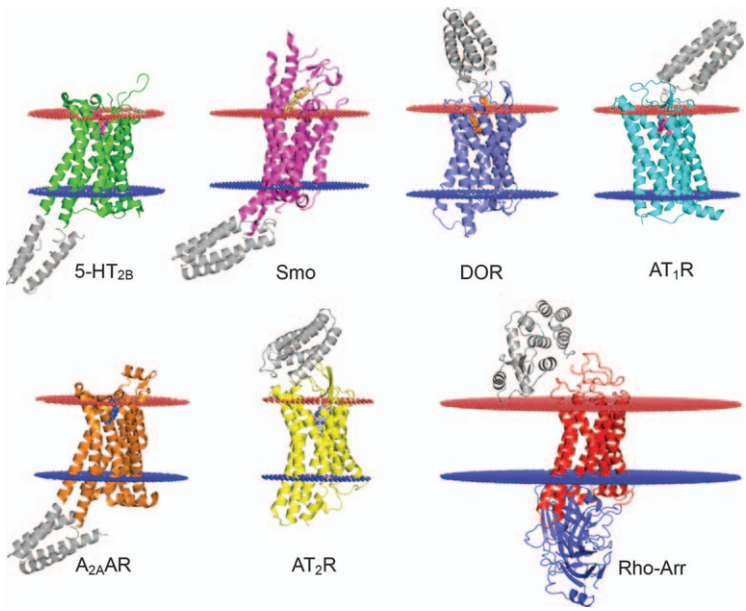


Figure 2.3 A gallery of GPCR structures determined by LCP-SFX. Serotonin 5-HT_{2B} receptor in complex with ergotamine (PDB ID 4NC3),[22] Smoothened receptor transmembrane domain in complex with cyclopamine (PDB ID 4O9R),[23] δ-opioid receptor in complex with a bi-functional peptide ligand DIPP (PDB ID 4RWD),[101] angiotensin II receptor type 1 in complex with ZD7155 (PDB ID 4YAY),[41] adenosine A_{2A} receptor in complex with ZM241385 (PDB ID 5K2C),[45] angiotensin II receptor type 2 in complex with a selective ligand compound 1 (PDB ID 5UNF, 5UNG),[42] rhodopsin–arrestin complex (PDB ID 4ZWJ).[10] Fusion proteins are shown in gray. Membrane boundaries (red and blue dots) are defined by the Orientation of Proteins in Membrane server (http://opm.phar.umich.edu).

Table 2.1 A summary of GPCR structures solved by LCP-SFX.

Receptor	Total data collection time (min)	Number of hits/ indexed images	Total protein used (mg)	Resolution (Å)	Space group	PDB ID	Ref.
5-HT_{2B}	590	152 651/32 819	0.3	2.8	$C222_1$	4NC3	22
Smo	490	274 214/61 964	0.5	3.2/3.4/4.0[a]	$P12_11$	4O9R	23
DOR	275	125 458/36 083	0.3	2.7	C121	4RWD	101
AT_1R	385	457 275/73 130	0.29	2.9	C121	4YAY	41
Rho–Arr	700	22 262/18 874	0.8	3.3/3.8/3.8[a]	$P2_12_12_1$	4ZWJ	10
$A_{2A}AR$	1020	1 797 503/578 620	2.7	1.9	$C222_1$	5K2C	45
	135	232 283/72 735	0.27				
AT_2R	375	175 241/22 774	0.19	2.8	$P12_11$	5UNF	42
		175 241/15 804		2.8	$P2_122_1$	5UNG	

[a]Anisotropic resolution cutoff.

room temperature using ~5 μm crystals was compared with the 2.7 Å resolution structure obtained by traditional microcrystallography at a synchrotron source using about 100 times larger by volume cryocooled crystals.[33] Both structures appeared mostly identical, except for the XFEL structure showing a distinct distribution of thermal motions and conformations of several residues that likely more accurately represent the receptor structure and dynamics in the native cellular environment.[22] After being properly validated, the LCP-SFX method was further applied to several GPCRs, structures of which at the time could not be determined using synchrotron beamlines, as briefly described below.

The smoothened (Smo) receptor belongs to the class frizzled GPCR and plays an essential role in the hedgehog signalling pathway during embryonal development and in carcinogenesis. While several Smo crystal structures in complex with small molecule antagonists and agonists were obtained at a synchrotron source,[32] the structure of Smo in complex with the naturally occuring teratogen cyclopamine resisted the efforts. Optimized crystals of Smo/cyclopamine with an average size of 120×10×5 μm produced poor diffraction at a microfocus synchrotron beamline showing high mosaicity, likely due to the accumulation of crystal growth defects or cryocooling. However, small (<5 μm) microcrystals at room temperature apparently did not suffer from this problem enabling structure determination by LCP-SFX at an anisotropic 3.4, 3.2 and 4.0 Å resolution along the principal crystal axes,[23] allowing for clear identification of cyclopamine in the binding pocket.

Morphine-based analgesics are widely prescribed to patients suffering from severe pain, however, their frequent consumption leads to acquired tolerance, dependence, and addiction. The function of such pain-killers is mostly mediated through the μ-opioid receptor. It has been shown that simultaneous activation of the μ-opioid receptor (MOR) along with blockade of the δ-opioid receptor (DOR) can prevent the development of morphine-induced tolerance and dependence, thus prompting for the design of such bifunctional ligands. In order to get insights in the binding mode, DOR was crystallized in complex with a bifunctional peptide DIPP-NH_2, and its initial structure was determined at a synchrotron source at 3.3 Å resolution. Due to the limited resolution, the placement of the peptide ligand in this structure was ambiguous, therefore a slightly different DOR construct was used to obtain ~5 μm in size microcrystals, which were used for LCP-SFX data collection, resulting in a 2.7 Å resolution structure with a clear electron density for the peptide ligand.[101] The DOR/DIPP-NH_2 structure provided a structural basis for better understanding of peptide binding and the development of new generation of bifunctional ligands.

Angiotensin II (AngII) is a peptide hormone, actions of which are mediated through two AngII receptors, AT_1R and AT_2R, sharing 34% of amino acid sequence identity. AT_1R is mostly responsible for the regulation of the blood pressure, while the function of AT_2R is more diverse and mysterious. Optimized AT_1R crystals in complex with an antagonist ZD7155 formed thin,

long needles which were difficult to harvest from the gel-like LCP without damaging them. These crystals diffracted to ~4.0 Å resolution at a synchrotron source. Similar needle-shaped but smaller crystals (~10×2×2 μm^3) yielded a 2.9 Å resolution structure at LCLS,[41] which became the first structure of a novel GPCR determined by LCP-SFX. The AT_1R/ZD7155 structure along with docking simulations of clinically used angiotensin receptor blockers revealed critical interactions providing fundamental insights in structure–function relationship and drug design. Following success with AT_1R, microcrystals (~5×2×2 μm^3) of AT_2R in complex with its selective ligand compound 1 (cpd1) were obtained in LCP and used for room temperature SFX data collection at LCLS.[42] Data processing revealed two distinct crystal forms, therefore resulting in two AT_2R/cpd1 structures, both at 2.8 Å resolution, from the same experiment. In parallel, AT_2R was crystallized with a dual AT_1R/AT_2R ligand, compound 2, forming larger crystals (~70×40×20 μm^3) that yielded a 2.9 Å resolution structure at a synchrotron source. In all these AT_2R structures the receptor was captured in an active-like conformation with helix VIII blocking the recruitment of G proteins and β-arrestin, in agreement with the lack of signaling responses in standard cellular assays. Structure–activity relationship (SAR), computer modeling, and mutagenesis studies revealed the crucial interactions that determine ligand binding specificity and selectivity. The structures therefore provide important insights into the structural basis of the distinct functions of the two AngII receptors and may guide design of new potent and selective ligands.

Finally, the structure of the visual rhodopsin in complex with its signaling partner arrestin represented another major milestone.[10] After extensive screening and optimizations, an engineered rhodopsin-arrestin (Rho–Arr) construct was crystallized in 9.7 MAG-based LCP supplemented with 10% cholesterol producing <20 μm crystals. Such crystals diffracted to only 7–8 Å resolution at a microfocus synchrotron beamline. Similar crystals of 5–10 μm in size diffracted substantially better at LCLS, and despite many challenges such as twinning, pseudosymmetry, and high anisotropy, the structure was solved to resolution limits of 3.3 Å, 3.8 Å and 3.8 Å along the principal crystal axes. The overall arrangement of the complex and specific receptor–arrestin interactions were validated by mutagenesis, disulfide cross-linking, hydrogen/deuterium exchange (HDX) mass spectrometery, double electron–electron resonance (DEER) spectrometry, and electron microscopy (EM). The structure of the complex provides fundamental insights into the general mechanism of arrestin-biased signaling and may contribute into development of functionally selective ligands.

The above examples clearly demonstrate the advantages of the LCP-SFX approach over traditional rotation/oscillation microcrystallography. GPCRs and other challenging-for-crystallization membrane proteins and complexes are uniquely matched for structural studies by this method because they often form well-ordered but small microcrystals in LCP, which are not suitable for high-resolution data collection at synchrotron beamlines due to accumulation of severe radiation damage. The LCP-SFX method has many

desirable aspects, such as the ability to obtain high-resolution structures from micron-sized crystals at room temperature without appreciable radiation damage, bypassing extensive crystal optimization, and crystal harvesting, while typically requiring ~0.3 mg of purified protein (Table 2.1). Due to the extremely short XFEL pulse duration and the transparency of LCP matrix, this approach also lends itself to ultrafast time-resolved studies of conformational changes in macromolecules triggered by a light laser pulse (pump) and probed by an XFEL pulse after specified time delay.[89] While the accessibility of the LCP-SFX method to the scientific community is currently limited by the small number of XFELs worldwide, it is expected that commissioning of new sources and further developments will produce a powerful impact on the structural biology of membrane proteins and other challenging macromolecules and their complexes.

Acknowledgements

This review is partly based on the research work in the author's laboratories supported by the National Institutes of Health grants R01 GM108635 (V.C.), P01 DA035764 (V.C.), R21 DA042298 (W.L.), the STC Program of the National Science Foundation through BioXFEL (No. 1231306) (U.W., W.L., V.C.), the Mayo Clinic-ASU Collaborative Seed Grant Award (W.L.) and the Flinn Foundation Seed Grant (W.L.).

References

1. J. S. Mason, A. Bortolato, M. Congreve and F. H. Marshall, *Trends Pharmacol. Sci.*, 2012, **33**, 249–260.
2. J. P. Overington, B. Al-Lazikani and A. L. Hopkins, *Nat. Rev. Drug Discovery*, 2006, **5**, 993–996.
3. E. M. Landau and J. P. Rosenbusch, *Proc. Natl. Acad. Sci. U. S. A.*, 1996, **93**, 14532–14535.
4. V. Cherezov, *Curr. Opin. Struct. Biol.*, 2011, **21**, 559–566.
5. V. Cherezov, M. A. Hanson, M. T. Griffith, M. C. Hilgart, R. Sanishvili, V. Nagarajan, S. Stepanov, R. F. Fischetti, P. Kuhn and R. C. Stevens, *J. R .Soc., Interface*, 2009, **6**(Suppl 5), S587–S597.
6. S. Boutet, L. Lomb, G. J. Williams, T. R. Barends, A. Aquila, R. B. Doak, U. Weierstall, D. P. Deponte, J. Steinbrener, R. L. Shoeman, M. Messerschmidt, A. Barty, T. A. White, S. Kassemeyer, R. A. Kirian, M. M. Seibert, P. A. Montanez, C. Kenney, R. Herbst, P. Hart, J. Pines, G. Haller, S. M. Gruner, H. T. Philipp, M. W. Tate, M. Hromalik, L. J. Koerner, N. van Bakel, J. Morse, W. Ghonsalves, D. Arnlund, M. J. Bogan, C. Caleman, R. Fromme, C. Y. Hampton, M. S. Hunter, L. Johansson, G. Katona, C. Kupitz, M. Liang, A. V. Martin, K. Nass, L. Redecke, F. Stellato, N. Timneanu, D. Wang, N. A. Zatsepin, D. Schafer, J. Defever, R. Neutze, P. Fromme, J. C. Spence, H. N. Chapman and I. Schlichting, *Science*, 2012, **337**, 362–364.

7. L. Redecke, K. Nass, D. P. DePonte, T. A. White, D. Rehders, A. Barty, F. Stellato, M. Liang, T. R. Barends, S. Boutet, G. J. Williams, M. Messerschmidt, M. M. Seibert, A. Aquila, D. Arnlund, S. Bajt, T. Barth, M. J. Bogan, C. Caleman, T. C. Chao, R. B. Doak, H. Fleckenstein, M. Frank, R. Fromme, L. Galli, I. Grotjohann, M. S. Hunter, L. C. Johansson, S. Kassemeyer, G. Katona, R. A. Kirian, R. Koopmann, C. Kupitz, L. Lomb, A. V. Martin, S. Mogk, R. Neutze, R. L. Shoeman, J. Steinbrener, N. Timneanu, D. Wang, U. Weierstall, N. A. Zatsepin, J. C. Spence, P. Fromme, I. Schlichting, M. Duszenko, C. Betzel and H. N. Chapman, *Science*, 2013, **339**, 227–230.
8. H. N. Chapman, P. Fromme, A. Barty, T. A. White, R. A. Kirian, A. Aquila, M. S. Hunter, J. Schulz, D. P. DePonte, U. Weierstall, R. B. Doak, F. R. Maia, A. V. Martin, I. Schlichting, L. Lomb, N. Coppola, R. L. Shoeman, S. W. Epp, R. Hartmann, D. Rolles, A. Rudenko, L. Foucar, N. Kimmel, G. Weidenspointner, P. Holl, M. Liang, M. Barthelmess, C. Caleman, S. Boutet, M. J. Bogan, J. Krzywinski, C. Bostedt, S. Bajt, L. Gumprecht, B. Rudek, B. Erk, C. Schmidt, A. Hömke, C. Reich, D. Pietschner, L. Strüder, G. Hauser, H. Gorke, J. Ullrich, S. Herrmann, G. Schaller, F. Schopper, H. Soltau, K. U. Kühnel, M. Messerschmidt, J. D. Bozek, S. P. Hau-Riege, M. Frank, C. Y. Hampton, R. G. Sierra, D. Starodub, G. J. Williams, J. Hajdu, N. Timneanu, M. M. Seibert, J. Andreasson, A. Rocker, O. Jönsson, M. Svenda, S. Stern, K. Nass, R. Andritschke, C. D. Schröter, F. Krasniqi, M. Bott, K. E. Schmidt, X. Wang, I. Grotjohann, J. M. Holton, T. R. Barends, R. Neutze, S. Marchesini, R. Fromme, S. Schorb, D. Rupp, M. Adolph, T. Gorkhover, I. Andersson, H. Hirsemann, G. Potdevin, H. Graafsma, B. Nilsson and J. C. Spence, *Nature*, 2011, **470**, 73–77.
9. L. C. Johansson, A. B. Wohri, G. Katona, S. Engstrom and R. Neutze, *Curr. Opin. Struct. Biol.*, 2009, **19**, 372–378.
10. Y. Kang, X. E. Zhou, X. Gao, Y. He, W. Liu, A. Ishchenko, A. Barty, T. A. White, O. Yefanov, G. W. Han, Q. Xu, P. W. de Waal, J. Ke, M. H. Tan, C. Zhang, A. Moeller, G. M. West, B. D. Pascal, N. Van Eps, L. N. Caro, S. A. Vishnivetskiy, R. J. Lee, K. M. Suino-Powell, X. Gu, K. Pal, J. Ma, X. Zhi, S. Boutet, G. J. Williams, M. Messerschmidt, C. Gati, N. A. Zatsepin, D. Wang, D. James, S. Basu, S. Roy-Chowdhury, C. E. Conrad, J. Coe, H. Liu, S. Lisova, C. Kupitz, I. Grotjohann, R. Fromme, Y. Jiang, M. Tan, H. Yang, J. Li, M. Wang, Z. Zheng, D. Li, N. Howe, Y. Zhao, J. Standfuss, K. Diederichs, Y. Dong, C. S. Potter, B. Carragher, M. Caffrey, H. Jiang, H. N. Chapman, J. C. Spence, P. Fromme, U. Weierstall, O. P. Ernst, V. Katritch, V. V. Gurevich, P. R. Griffin, W. L. Hubbel, R. C. Stevens, V. Cherezov, K. Melcher and H. E. Xu, *Nature*, 2015, **523**, 561–567.
11. V. Cherezov, D. M. Rosenbaum, M. A. Hanson, S. G. Rasmussen, F. S. Thian, T. S. Kobilka, H. J. Choi, P. Kuhn, W. I. Weis, B. K. Kobilka and R. C. Stevens, *Science*, 2007, **318**, 1258–1265.

12. D. M. Rosenbaum, V. Cherezov, M. A. Hanson, S. G. F. Rasmussen, S. T. Foon, T. S. Kobilka, H.-J. Choi, X.-J. Yao, W. I. Weis, R. C. Stevens and B. K. Kobilka, *Science*, 2007, **318**, 1266–1273.
13. V. P. Jaakola, M. T. Griffith, M. A. Hanson, V. Cherezov, E. Y. Chien, J. R. Lane, A. P. Ijzerman and R. C. Stevens, *Science*, 2008, **322**, 1211–1217.
14. A. I. Alexandrov, M. Mileni, E. Y. Chien, M. A. Hanson and R. C. Stevens, *Structure*, 2008, **16**, 351–359.
15. M. Caffrey and V. Cherezov, *Nat. Protoc.*, 2009, **4**, 706–731.
16. V. Cherezov, E. Abola and R. C. Stevens, *Methods Mol. Biol.*, 2010, **654**, 141–168.
17. W. Liu, M. A. Hanson, R. C. Stevens and V. Cherezov, *Biophys. J.*, 2010, **98**, 1539–1548.
18. W. Liu and V. Cherezov, *J. Visualized Exp.*, 2011, **49**, e2501.
19. F. Xu, W. Liu, M. A. Hanson, R. C. Stevens and V. Cherezov, *Cryst. Growth Des.*, 2011, **11**, 1193–1201.
20. E. Chun, A. A. Thompson, W. Liu, C. B. Roth, M. T. Griffith, V. Katritch, J. Kunken, F. Xu, V. Cherezov, M. A. Hanson and R. C. Stevens, *Structure*, 2012, **20**, 967–976.
21. R. C. Stevens, V. Cherezov, V. Katritch, R. Abagyan, P. Kuhn, H. Rosen and K. Wuthrich, *Nat. Rev. Drug Discovery*, 2013, **12**, 25–34.
22. W. Liu, D. Wacker, C. Gati, G. W. Han, D. James, D. Wang, G. Nelson, U. Weierstall, V. Katritch, A. Barty, N. A. Zatsepin, D. Li, M. Messerschmidt, S. Boutet, G. J. Williams, J. E. Koglin, M. M. Seibert, C. Wang, S. T. Shah, S. Basu, R. Fromme, C. Kupitz, K. N. Rendek, I. Grotjohann, P. Fromme, R. A. Kirian, K. R. Beyerlein, T. A. White, H. N. Chapman, M. Caffrey, J. C. Spence, R. C. Stevens and V. Cherezov, *Science*, 2013, **342**, 1521–1524.
23. U. Weierstall, D. James, C. Wang, T. A. White, D. Wang, W. Liu, J. C. Spence, R. Bruce Doak, G. Nelson, P. Fromme, R. Fromme, I. Grotjohann, C. Kupitz, N. A. Zatsepin, H. Liu, S. Basu, D. Wacker, G. W. Han, V. Katritch, S. Boutet, M. Messerschmidt, G. J. Williams, J. E. Koglin, M. Marvin Seibert, M. Klinker, C. Gati, R. L. Shoeman, A. Barty, H. N. Chapman, R. A. Kirian, K. R. Beyerlein, R. C. Stevens, D. Li, S. T. Shah, N. Howe, M. Caffrey and V. Cherezov, *Nat. Commun.*, 2014, **5**, 3309.
24. W. Liu, A. Ishchenko and V. Cherezov, *Nat. Protoc.*, 2014, **9**, 2123–2134.
25. E. Y. Chien, W. Liu, Q. Zhao, V. Katritch, G. W. Han, M. A. Hanson, L. Shi, A. H. Newman, J. A. Javitch, V. Cherezov and R. C. Stevens, *Science*, 2010, **330**, 1091–1095.
26. B. Wu, E. Y. Chien, C. D. Mol, G. Fenalti, W. Liu, V. Katritch, R. Abagyan, A. Brooun, P. Wells, F. C. Bi, D. J. Hamel, P. Kuhn, T. M. Handel, V. Cherezov and R. C. Stevens, *Science*, 2010, **330**, 1066–1071.
27. T. Shimamura, M. Shiroishi, S. Weyand, H. Tsujimoto, G. Winter, V. Katritch, R. Abagyan, V. Cherezov, W. Liu, G. W. Han, T. Kobayashi, R. C. Stevens and S. Iwata, *Nature*, 2011, **475**, 65–72.

28. H. Wu, D. Wacker, M. Mileni, V. Katritch, G. W. Han, E. Vardy, W. Liu, A. A. Thompson, X. P. Huang, F. I. Carroll, S. W. Mascarella, R. B. Westkaemper, P. D. Mosier, B. L. Roth, V. Cherezov and R. C. Stevens, *Nature*, 2012, **485**, 327–332.
29. A. A. Thompson, W. Liu, E. Chun, V. Katritch, H. X. Wu, E. Vardy, X. P. Huang, C. Trapella, R. Guerrini, G. Calo, B. L. Roth, V. Cherezov and R. C. Stevens, *Nature*, 2012, **485**, 395–399.
30. M. A. Hanson, C. B. Roth, E. Jo, M. T. Griffith, F. L. Scott, G. Reinhart, H. Desale, B. Clemons, S. M. Cahalan, S. C. Schuerer, M. G. Sanna, G. W. Han, P. Kuhn, H. Rosen and R. C. Stevens, *Science*, 2012, **335**, 851–855.
31. C. Wang, Y. Jiang, J. Ma, H. Wu, D. Wacker, V. Katritch, G. W. Han, W. Liu, X. P. Huang, E. Vardy, J. D. McCorvy, X. Gao, X. E. Zhou, K. Melcher, C. Zhang, F. Bai, H. Yang, L. Yang, H. Jiang, B. L. Roth, V. Cherezov, R. C. Stevens and H. E. Xu, *Science*, 2013, **340**, 610–614.
32. C. Wang, H. Wu, V. Katritch, G. W. Han, X.-P. Huang, W. Liu, F. Y. Siu, B. L. Roth, V. Cherezov and R. C. Stevens, *Nature*, 2013, **497**, 338–343.
33. D. Wacker, C. Wang, V. Katritch, G. W. Han, X. P. Huang, E. Vardy, J. D. McCorvy, Y. Jiang, M. Chu, F. Y. Siu, W. Liu, H. E. Xu, V. Cherezov, B. L. Roth and R. C. Stevens, *Science*, 2013, **340**, 615–619.
34. F. Y. Siu, M. He, C. de Graaf, G. W. Han, D. Yang, Z. Zhang, C. Zhou, Q. Xu, D. Wacker, J. S. Joseph, W. Liu, J. Lau, V. Cherezov, V. Katritch, M. W. Wang and R. C. Stevens, *Nature*, 2013, **499**, 444–449.
35. Q. Tan, Y. Zhu, J. Li, Z. Chen, G. W. Han, I. Kufareva, T. Li, L. Ma, G. Fenalti, W. Zhang, X. Xie, H. Yang, H. Jiang, V. Cherezov, H. Liu, R. C. Stevens, Q. Zhao and B. Wu, *Science*, 2013, **341**, 1387–1390.
36. H. Wu, C. Wang, K. J. Gregory, G. W. Han, H. P. Cho, Y. Xia, C. M. Niswender, V. Katritch, J. Meiler, V. Cherezov, P. J. Conn and R. C. Stevens, *Science*, 2014, **344**, 58–64.
37. K. Zhang, J. Zhang, Z.-G. Gao, D. Zhang, L. Zhu, G. W. Han, S. M. Moss, S. Paoletta, E. Kiselev, W. Lu, G. Fenalti, W. Zhang, C. E. Muller, H. Yang, H. Jiang, V. Cherezov, V. Katritch, K. A. Jacobson, R. C. Stevens, B. Wu and Q. Zhao, *Nature*, 2014, **509**, 115–118.
38. G. Fenalti, P. M. Giguere, V. Katritch, X.-P. Huang, A. A. Thompson, V. Cherezov, B. L. Roth and R. C. Stevens, *Nature*, 2014, **506**, 191–196.
39. D. Zhang, Z.-G. Gao, K. Zhang, E. Kiselev, S. Crane, J. Wang, S. Paoletta, C. Yi, L. Ma, W. Zhang, G. W. Han, H. Liu, V. Cherezov, V. Katritch, H. Jiang, R. C. Stevens, K. A. Jacobson, Q. Zhao and B. Wu, *Nature*, 2015, **520**, 317.
40. J. E. Chrencik, C. B. Roth, M. Terakado, H. Kurata, R. Omi, Y. Kihara, D. Warshaviak, S. Nakade, G. Asmar-Rovira, M. Mileni, H. Mizuno, M. T. Griffith, C. Rodgers, G. W. Han, J. Velasquez, J. Chun, R. C. Stevens and M. A. Hanson, *Cell*, 2015, **161**, 1633–1643.
41. H. Zhang, H. Unal, C. Gati, G. W. Han, W. Liu, N. A. Zatsepin, D. James, D. Wang, G. Nelson, U. Weierstall, M. R. Sawaya, Q. Xu, M. Messerschmidt, G. J. Williams, S. Boutet, O. M. Yefanov, T. A. White,

C. Wang, A. Ishchenko, K. C. Tirupula, R. Desnoyer, J. Coe, C. E. Conrad, P. Fromme, R. C. Stevens, V. Katritch, S. S. Karnik and V. Cherezov, *Cell*, 2015, **161**, 833–844.
42. H. Zhang, G. W. Han, A. Batyuk, A. Ishchenko, K. L. White, N. Patel, A. Sadybekov, B. Zamlynny, M. T. Rudd, K. Hollenstein, A. Tolstikova, T. A. White, M. S. Hunter, U. Weierstall, W. Liu, K. Babaoglu, E. L. Moore, R. D. Katz, J. M. Shipman, M. Garcia-Calvo, S. Sharma, P. Sheth, S. M. Soisson, R. C. Stevens, V. Katritch and V. Cherezov, *Nature*, 2017, DOI: 10.1038/nature22035.
43. J. Steyaert and B. K. Kobilka, *Curr. Opin. Struct. Biol.*, 2011, **21**, 567–572.
44. T. A. White, R. A. Kirian, A. V. Martin, A. Aquila, K. Nass, A. Barty and H. N. Chapman, J. Appl. Crystallo*gr.*, 2012, **45**, 335–341.
45. A. Batyuk, L. Galli, A. Ishchenko, G. W. Han, C. Gati, P. A. Popov, M. Y. Lee, B. Stauch, T. A. White, A. Barty, A. Aquila, M. S. Hunter, M. Liang, S. Boutet, M. Pu, Z. J. Liu, G. Nelson, D. James, C. Li, Y. Zhao, J. C. Spence, W. Liu, P. Fromme, V. Katritch, U. Weierstall, R. C. Stevens and V. Cherezov, *Sci. Adv.*, 2016, **2**, e1600292.
46. K. Yamashita, D. Pan, T. Okuda, M. Sugahara, A. Kodan, T. Yamaguchi, T. Murai, K. Gomi, N. Kajiyama, E. Mizohata, M. Suzuki, E. Nango, K. Tono, Y. Joti, T. Kameshima, J. Park, C. Song, T. Hatsui, M. Yabashi, S. Iwata, H. Kato, H. Ago, M. Yamamoto and T. Nakatsu, *Sci. Rep.*, 2015, **5**, 14017.
47. J. Xiang, E. Chun, C. Liu, L. Jing, Z. Al-Sahouri, L. Zhu and W. Liu, *Trends Pharmacol. Sci.*, 2016, **37**, 1055–1069.
48. E. Ghosh, P. Kumari, D. Jaiman and A. K. Shukla, *Nat. Rev. Mol. Cell Biol.*, 2015, **16**, 69–81.
49. R. A. Abagyan, A. Orry, E. Raush, L. Budagyan and M. Totrov, *ICM*, Molsoft LLC, La Jolla, CA, 2015.
50. S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess and E. Lindahl, *Bioinformatics*, 2013, **29**, 845–854.
51. M. M. Ghahremanpour, S. S. Arab, S. B. Aghazadeh, J. Zhang and D. van der Spoel, *Bioinformatics*, 2014, **30**, 439–441.
52. E. Segala, D. Guo, R. K. Cheng, A. Bortolato, F. Deflorian, A. S. Doré, J. C. Errey, L. H. Heitman, A. P. IJzerman, F. H. Marshall and R. M. Cooke, *J. Med. Chem.*, 2016, **59**, 6470–6479.
53. W. Liu, E. Chun, A. A. Thompson, P. Chubukov, F. Xu, V. Katritch, G. W. Han, C. B. Roth, L. H. Heitman, A. P. IJzerman, V. Cherezov and R. C. Stevens, *Science*, 2012, **337**, 232–236.
54. Z. Shao, J. Yin, K. Chapman, M. Grzemska, L. Clark, J. Wang and D. M. Rosenbaum, *Nature*, 2016, **540**, 602–606.
55. J. Yin, J. C. Mobarec, P. Kolb and D. M. Rosenbaum, *Nature*, 2015, **519**, 247–250.
56. J. Yin, K. Babaoglu, C. A. Brautigam, L. Clark, Z. Shao, T. H. Scheuermann, C. M. Harrell, A. L. Gotter, A. J. Roecker,

C. J. Winrow, J. J. Renger, P. J. Coleman and D. M. Rosenbaum, *Nat. Struct. Mol. Biol.*, 2016, **23**, 293–299.
57. K. Hollenstein, J. Kean, A. Bortolato, R. K. Cheng, A. S. Doré, A. Jazayeri, R. M. Cooke, M. Weir and F. H. Marshall, *Nature*, 2013, **499**, 438–443.
58. A. Jazayeri, A. S. Doré, D. Lamb, H. Krishnamurthy, S. M. Southall, A. H. Baig, A. Bortolato, M. Koglin, N. J. Robertson, J. C. Errey, S. P. Andrews, I. Teobald, A. J. Brown, R. M. Cooke, M. Weir and F. H. Marshall, *Nature*, 2016, **533**, 274–277.
59. A. S. Doré, K. Okrasa, J. C. Patel, M. Serrano-Vega, K. Bennett, R. M. Cooke, J. C. Errey, A. Jazayeri, S. Khan, B. Tehan, M. Weir, G. R. Wiggin and F. H. Marshall, *Nature*, 2014, **511**, 557–562.
60. P. Egloff, M. Hillenbrand, C. Klenk, A. Batyuk, P. Heine, S. Balada, K. M. Schlinkmann, D. J. Scott, M. Schutz and A. Pluckthun, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, E655–E662.
61. J. S. Burg, J. R. Ingram, A. J. Venkatakrishnan, K. M. Jude, A. Dukkipati, E. N. Feinberg, A. Angelini, D. Waghray, R. O. Dror, H. L. Ploegh and K. C. Garcia, *Science*, 2015, **347**, 1113–1117.
62. T. Hua, K. Vemuri, M. Pu, L. Qu, G. W. Han, Y. Wu, S. Zhao, W. Shui, S. Li, A. Korde, R. B. Laprairie, E. L. Stahl, J. H. Ho, N. Zvonok, H. Zhou, I. Kufareva, B. Wu, Q. Zhao, M. A. Hanson, L. M. Bohn, A. Makriyannis, R. C. Stevens and Z. J. Liu, *Cell*, 2016, **167**, 750–762 e714.
63. M. A. Hanson, A. Brooun, K. A. Baker, V. P. Jaakola, C. Roth, E. Y. Chien, A. Alexandrov, J. Velasquez, L. Davis, M. Griffith, K. Moy, B. K. Ganser-Pornillos, Y. Hua, P. Kuhn, S. Ellis, M. Yeager and R. C. Stevens, *Protein Expression Purif.*, 2007, **56**, 85–92.
64. C. B. Roth, M. A. Hanson and R. C. Stevens, *J. Mol. Biol.*, 2008, **376**, 1305–1319.
65. T. T. Leskela, P. M. Markkanen, E. M. Pietila, J. T. Tuusa and U. E. Petaja-Repo, *J. Biol. Chem.*, 2007, **282**, 23171–23183.
66. J. P. Morello, A. Salahpour, A. Laperriere, V. Bernier, M. F. Arthus, M. Lonergan, U. Petaja-Repo, S. Angers, D. Morin, D. G. Bichet and M. Bouvier, *J. Clin. Invest.*, 2000, **105**, 887–895.
67. G. Gimpl and F. Fahrenholz, *Biochim. Biophys. Acta*, 2002, **1564**, 384–392.
68. M. A. Hanson, V. Cherezov, M. T. Griffith, C. B. Roth, V. P. Jaakola, E. Y. Chien, J. Velasquez, P. Kuhn and R. C. Stevens, *Structure*, 2008, **16**, 897–905.
69. T. J. Pucadyil and A. Chattopadhyay, *Biochim. Biophys. Acta*, 2004, **1663**, 188–200.
70. T. J. Pucadyil and A. Chattopadhyay, *J. Fluoresc.*, 2006, **16**, 87–94.
71. B. L. Sprague and J. G. McNally, *Trends Cell Biol.*, 2005, **15**, 84–91.
72. V. Cherezov, J. Liu, M. Griffith, M. A. Hanson and R. C. Stevens, *Cryst. Growth Des.*, 2008, **8**, 4307–4315.
73. H. Zhang, H. Unal, R. Desnoyer, G. W. Han, N. Patel, V. Katritch, S. S. Karnik, V. Cherezov and R. C. Stevens, *J. Biol. Chem.*, 2015, **290**, 29127–29139.

74. G. G. Prive, *Methods*, 2007, **41**, 388–397.
75. A. A. Thompson, J. J. Liu, E. Chun, D. Wacker, H. Wu, V. Cherezov and R. C. Stevens, *Methods*, 2011, **55**, 310–317.
76. P. S. Chae, S. G. Rasmussen, R. R. Rana, K. Gotfryd, R. Chandra, M. A. Goren, A. C. Kruse, S. Nurva, C. J. Loland, Y. Pierre, D. Drew, J. L. Popot, D. Picot, B. G. Fox, L. Guan, U. Gether, B. Byrne, B. Kobilka and S. H. Gellman, *Nat. Methods*, 2010, **7**, 1003–1008.
77. C. V. Kulkarni, W. Wachter, G. Iglesias-Salto, S. Engelskirchen and S. Ahualli, *Phys. Chem. Chem. Phys.*, 2011, **13**, 3004–3021.
78. V. Cherezov, A. Peddi, L. Muthusubramaniam, Y. F. Zheng and M. Caffrey, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2004, **60**, 1795–1807.
79. D. Li, C. Boland, K. Walsh and M. Caffrey, *J. Visualized Exp.*, 2012, e4000.
80. A. Ishchenko, V. Cherezov and W. Liu, *Biophys. J.*, 2016, **110**, 59A.
81. L. V. Misquitta, Y. Misquitta, V. Cherezov, O. Slattery, J. M. Mohan, D. Hart, M. Zhalnina, W. A. Cramer and M. Caffrey, *Structure*, 2004, **12**, 2113–2124.
82. Y. Misquitta, V. Cherezov, F. Havas, S. Patterson, J. M. Mohan, A. J. Wells, D. J. Hart and M. Caffrey, *J. Struct. Biol.*, 2004, **148**, 169–175.
83. H. Qiu and M. Caffrey, *Biomaterials*, 2000, **21**, 223–234.
84. V. Cherezov, H. Fersi and M. Caffrey, *Biophys. J.*, 2001, **81**, 225–242.
85. J. S. Joseph, W. Liu, J. Kunken, T. M. Weiss, H. Tsuruta and V. Cherezov, *Methods*, 2011, **55**, 342–349.
86. C. Bostedt, S. Boutet, D. M. Fritz, Z. Huang, H. J. Lee, H. T. Lemke, A. Robert, W. F. Schlotter, J. J. Turner and G. J. Williams, *Rev. Mod. Phys.*, 2016, **88**, 015007.
87. M. Liang, G. J. Williams, M. Messerschmidt, M. M. Seibert, P. A. Montanez, M. Hayes, D. Milathianaki, A. Aquila, M. S. Hunter, J. E. Koglin, D. W. Schafer, S. Guillet, A. Busse, R. Bergan, W. Olson, K. Fox, N. Stewart, R. Curtis, A. A. Miahnahri and S. Boutet, *J. Synchrotron Radiat.*, 2015, **22**, 514–519.
88. G. Blaj, P. Caragiulo, G. Carini, S. Carron, A. Dragone, D. Freytag, G. Haller, P. Hart, J. Hasi, R. Herbst, S. Herrmann, C. Kenney, B. Markovic, K. Nishimura, S. Osier, J. Pines, B. Reese, J. Segal, A. Tomada and M. Weaver, *J. Synchrotron Radiat.*, 2015, **22**, 577–583.
89. Y. Fukuda, K. M. Tse, T. Nakane, T. Nakatsu, M. Suzuki, M. Sugahara, S. Inoue, T. Masuda, F. Yumoto, N. Matsugaki, E. Nango, K. Tono, Y. Joti, T. Kameshima, C. Song, T. Hatsui, M. Yabashi, O. Nureki, M. E. Murphy, T. Inoue, S. Iwata and E. Mizohata, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, 2928–2933.
90. A. Barty, R. A. Kirian, F. R. Maia, M. Hantke, C. H. Yoon, T. A. White and H. Chapman, *J. Appl. Crystallogr.*, 2014, **47**, 1118–1131.
91. J. Hattne, N. Echols, R. Tran, J. Kern, R. J. Gildea, A. S. Brewster, R. Alonso-Mori, C. Glöckner, J. Hellmich, H. Laksmono, R. G. Sierra, B. Lassalle-Kaiser, A. Lampe, G. Han, S. Gul, D. DiFiore,

D. Milathianaki, A. R. Fry, A. Miahnahri, W. E. White, D. W. Schafer, M. M. Seibert, J. E. Koglin, D. Sokaras, T. C. Weng, J. Sellberg, M. J. Latimer, P. Glatzel, P. H. Zwart, R. W. Grosse-Kunstleve, M. J. Bogan, M. Messerschmidt, G. J. Williams, S. Boutet, J. Messinger, A. Zouni, J. Yano, U. Bergmann, V. K. Yachandra, P. D. Adams and N. K. Sauter, *Nat. Methods*, 2014, **11**, 545–548.

92. W. Kabsch, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2014, **70**, 2204–2216.
93. W. Brehm and K. Diederichs, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2014, **70**, 101–109.
94. T. A. White, V. Mariani, W. Brehm, O. Yefanov, A. Barty, K. R. Beyerlein, F. Chervinskii, L. Galli, C. Gati, T. Nakane, A. Tolstikova, K. Yamashita, C. H. Yoon, K. Diederichs and H. N. Chapman, *J. Appl. Crystallogr.*, 2016, **49**, 680–689.
95. R. A. Kirian, X. Wang, U. Weierstall, K. E. Schmidt, J. C. Spence, M. Hunter, P. Fromme, T. White, H. N. Chapman and J. Holton, *Opt. Express*, 2010, **18**, 5713–5723.
96. O. Yefanov, V. Mariani, C. Gati, T. A. White, H. N. Chapman and A. Barty, *Opt. Express*, 2015, **23**, 28459–28470.
97. M. Uervirojnangkoorn, O. B. Zeldin, A. Y. Lyubimov, J. Hattne, A. S. Brewster, N. K. Sauter, A. T. Brunger and W. I. Weis, *eLife*, 2015, **4**, e05421.
98. H. M. Ginn, M. Messerschmidt, X. Ji, H. Zhang, D. Axford, R. J. Gildea, G. Winter, A. S. Brewster, J. Hattne, A. Wagner, J. M. Grimes, G. Evans, N. K. Sauter, G. Sutton and D. I. Stuart, *Nat. Commun.*, 2015, **6**, 6435.
99. T. A. White, *Philos. Trans. R. Soc. London*, 2014, **369**, 20130330.
100. H. M. Ginn, A. S. Brewster, J. Hattne, G. Evans, A. Wagner, J. M. Grimes, N. K. Sauter, G. Sutton and D. I. Stuart, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2015, **71**, 1400–1410.
101. G. Fenalti, N. A. Zatsepin, C. Betti, P. Giguere, G. W. Han, A. Ishchenko, W. Liu, K. Guillemyn, H. Zhang, D. James, D. Wang, U. Weierstall, J. C. Spence, S. Boutet, M. Messerschmidt, G. J. Williams, C. Gati, O. M. Yefanov, T. A. White, D. Oberthuer, M. Metz, C. H. Yoon, A. Barty, H. N. Chapman, S. Basu, J. Coe, C. E. Conrad, R. Fromme, P. Fromme, D. Tourwe, P. W. Schiller, B. L. Roth, S. Ballet, V. Katritch, R. C. Stevens and V. Cherezov, *Nat. Struct. Mol. Biol.*, 2015, **22**, 265–268.

CHAPTER 3

The Mesh&Collect Pipeline for the Collection of Multi-crystal Data Sets in Macromolecular Crystallography

NICOLAS FOOS,[a] GLEB BOURENKOV,[b] GORDON LEONARD,*[a] IGOR MELNIKOV,[a] CHRISTOPH MUELLER-DIECKMANN,[a] MAX NANAO,[a] ALEXANDER POPOV,[a] GIANLUCA SANTONI[a] AND ULRICH ZANDER[a,†]

[a] Structural Biology Group, European Synchrotron Radiation Facility, CS 40220, Grenoble 38043, France; [b] European Molecular Biology Laboratory (EMBL), Hamburg Outstation, Notkestrasse 85, 22607 Hamburg, Germany
*Email: leonard@esrf.fr

3.1 Introduction

The advent of cryo-cooling as a generally accepted technique[1–3] in macromolecular crystallography (MX) has meant that MX has, essentially, become a 'single crystal method' in which, while care has to be taken not to induce too much radiation damage to a sample,[4,5] the vast majority of data sets used for structure solution and refinement are collected from a single crystal. However, and as is often not understood by those new to MX, the resolution

[†]Current address: European Molecular Biology Laboratory, Grenoble Outstation, 71 avenue des Martyrs, CS 90181, 38042 Grenoble Cedex 9, France.

Chemical Biology No. 8
Protein Crystallography: Challenges and Practical Solutions
Edited by Konstantinos Beis and Gwyndaf Evans
© The Royal Society of Chemistry 2018
Published by the Royal Society of Chemistry, www.rsc.org

of a data set obtainable from a single crystal before it is destroyed by radiation damage *is* limited.[6] Moreover, the resolution of a complete data set obtainable is a function of crystal volume. Indeed, and as is shown in Figure 3.1, even for robust, well-ordered crystals, the smaller the sample the lower the resolution of the complete data set that is obtainable once radiation damage is taken into account. Thus, if experimenters are to obtain a data set to the maximum possible resolution from small crystals of a particular biological macromolecule, this has to be constructed by merging partial data sets from more than one crystal.

Such 'multi-crystal' data collection is not a new idea and goes back to the very early days of MX where, for example, the solution of the crystal structure of myoglobin was carried out using the Isomorphous Replacement (IR) technique, with data sets for each of the heavy atom derivatives used compiled from partial data sets (in this case single oscillation images) from as many as 22 different crystals.[7] While the collection, processing and merging of partial data sets from 22 different crystals is onerous for the experimenter, the effort involved is not insurmountable. However, the ultra-brilliant X-ray beams now available at X-ray free-electron lasers (XFELs) and 3rd generation synchrotron sources mean that measurable diffraction data can be obtained from *very* small crystals, often less than one μm in maximum dimension. Here many thousands of partial data sets are required to obtain a complete data set to the maximum possible resolution (Figure 3.1). The collection, processing and merging of these would be almost impossible in the absence of new paradigms.

The most spectacular new paradigm – serial femtosecond crystallography (SFX) – for the collection and processing of partial data sets from several thousand (and more) crystals of the same sample, was developed at XFELs.[8]

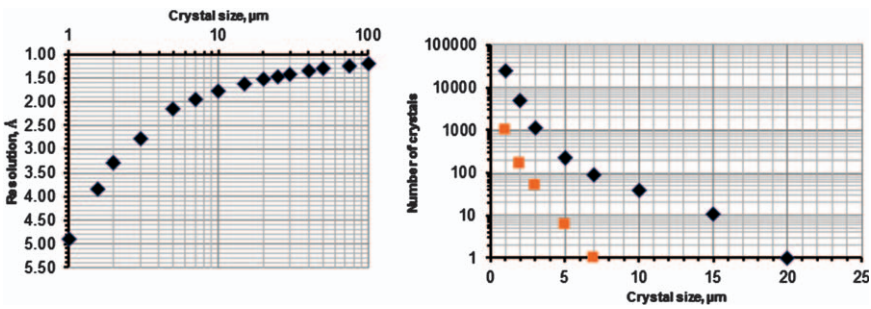


Figure 3.1 Complete data set resolution achievable from cryo-cooled single crystals taking a Henderson/Garman limit absorbed dose during data collection into account. Left: resolution *vs.* crystal size. Right: Number of crystals of a given size required to achieve complete data set resolutions of 1.5 Å (black) and 2.0 Å (red). Examples shown are for crystals of thermolysin, space group $P6_122$; Wilson B-factor = 11.5 $Å^2$. For a crystal 1 $μm^3$ in dimensions partial data sets from about 1000 crystals would be needed to achieve a final data set resolution of 2.0 Å.

Here, microcrystals are streamed through the X-ray beam, 'still' diffraction images are collected when the crystal and XFEL X-ray laser pulse coincide and complete data sets are then compiled by merging diffraction spot intensities from many thousands of different images. SFX will undoubtedly prove a seminal technique in MX, and similar experiments (SSX – serial synchrotron crystallography) are now being carried out at synchrotron sources.[9–11] However, such experiments are currently not without their disadvantages. In particular, the diffraction images collected contain, in the main, partially recorded reflections measured from crystals of different sizes and with (at XFELs) laser pulses of different spectral content so data quality does not yet approach that obtainable in single crystal MX experiments, where the continuous oscillation of the crystal in the X-ray beam allows a much more robust estimate of the intensity of a given reflection and its standard deviation. Moreover, the majority of images collected in such experiments contain no useful diffraction as synchronisation of crystal arrival, X-ray pulse (in the case of XFELs) and detector readout is not straightforward. This latter aspect also means that the amount of protein required in such streaming-based experiments is currently rather high even for structure solution based on molecular-replacement techniques (*i.e.* only native data sets are collected).

To try to overcome some of these disadvantages, Gati *et al.*, 2014[12] developed a variant of SSX in which the sample was contained on a cryo-cooled sample holder which was then continuously rastered through the X-ray beam while at the same time being rotated. Diffraction images were then recorded on-the-fly at set time intervals. As for experiments in which crystals are introduced into the X-ray beam in liquid streams, most of the diffraction images collected contained no useful information as crystal position, X-ray beam and detector readout did not coincide. However, when they did coincide, that the sample was being both translated and rotated meant that the diffraction images did not contain only partial reflections. This made the processing and scaling of the diffraction data less complicated, improving overall data quality. Nevertheless, this way of collecting serial data sets is not optimal, particularly as no attempt was made to synchronize the intersection of the X-ray beam and crystal with detector readout. Moreover, the absorbed X-ray dose delivered during the collection of each single diffraction image was not optimised to take advantage of the Henderson/Garman limits that cryo-cooled crystals of biological macromolecules can sustain.

Clearly, some sort of pre-interrogation of the sample holder,[14,15] followed by data collection from each crystal in turn, would both improve sample consumption and improve data quality. However, both pre-interrogation of the sample and the collection and merging of the many partial data sets collected from the crystals identified would be onerous if carried out manually. To further optimise and automate the collection of diffraction data in serial crystallography experiments from samples from which many X-ray diffraction images can be obtained before significant radiation

damage occurs, we developed the Mesh&Collect (Figure 3.2) protocol.[13] In this protocol, the positions of many randomly oriented (micro)crystals contained in a single sample holder are determined using a low dose X-ray-based mesh scan of the sample holder, the diffraction strength of each crystal found is determined and ranked and partial data sets are collected from as many crystals as the experimenter desires. Each partial data set is processed on-line.[16,17] Hierarchical cluster analysis (HCA) is then employed to choose which partial data sets to merge to produce a final, complete data set for use in downstream structure solution and refinement. Mesh&Collect can, in principle, be applied to crystals mounted in almost any type of rigid mounting platform including nylon loops, micro-meshes, Si_3N_4[18] and COC

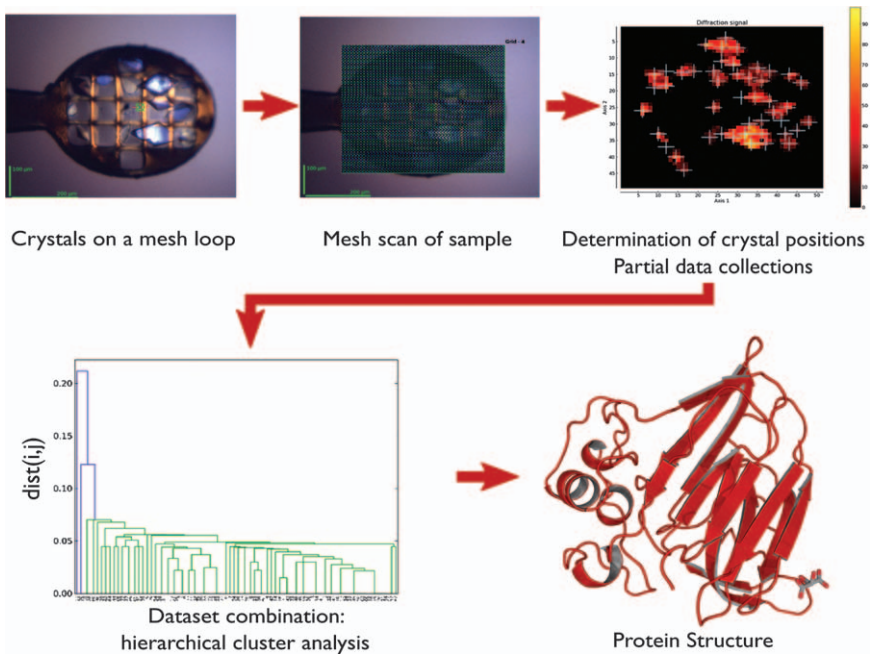
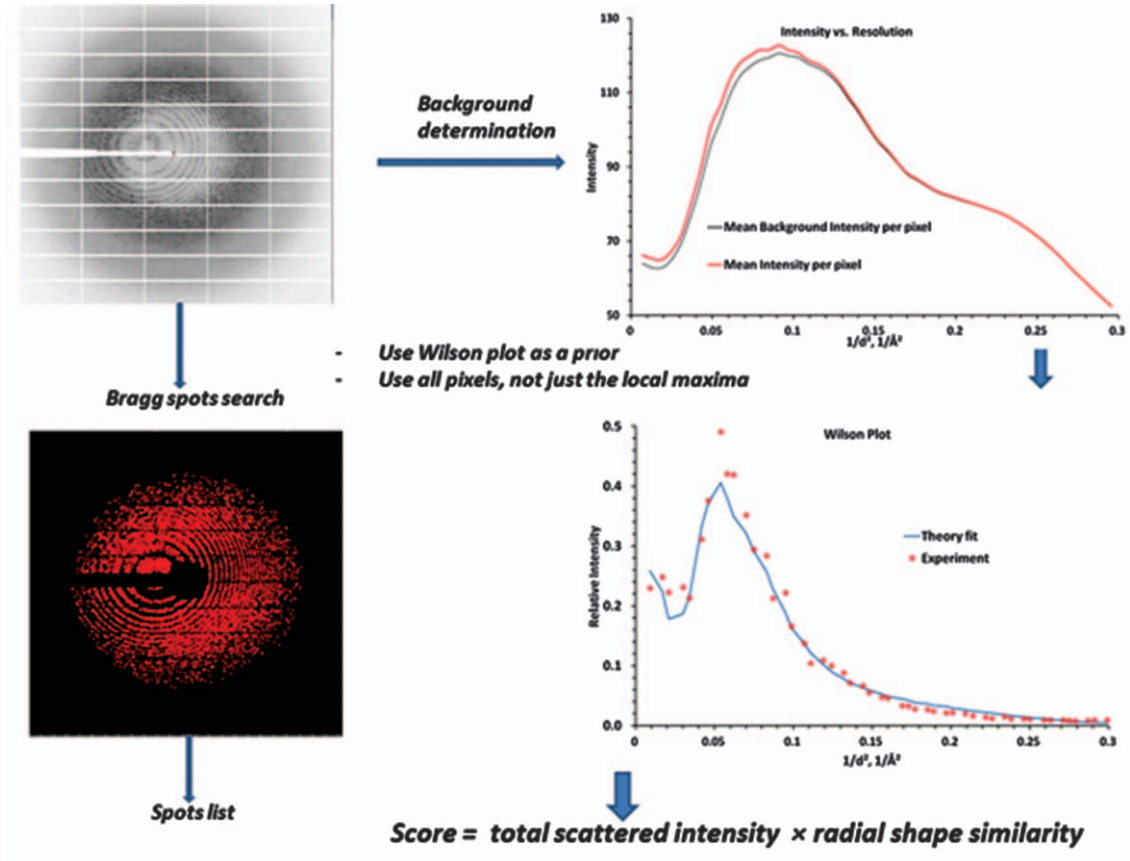


Figure 3.2 MeshAndCollect for multi-crystal data-collection. From top left: Many crystals from the same crystallisation drop are mounted on the same sample support; a pre-interrogation of the sample holder using a X-ray mesh scan is carried out and the images obtained are automatically analysed for diffraction signal and ranked according to diffraction strength. This produces a map on which positions for potential partial data set collection are shown. Once the experimenter has chosen the desired data collection parameters, a data-collection queue is automatically filled and all partial data sets are collected and automatically processed. Hierarchical cluster analysis (HCA) is then used to guide the choice of which partial data sets to merge in order to produce the final data set for use in downstream structure solution and/or refinement processes.
Adapted from Zander *et al.*, 2015[13] and reproduced with permission of the International Union of Crystallography (http://journals.iucr.org/).


Background determination
Intensity vs. Resolution
Intensity
Mean Background Intensity per pixel
Mean Intensity per pixel
1/d², 1/Å²
Bragg spots search
- Use Wilson plot as a prior
- Use all pixels, not just the local maxima
Wilson Plot
Relative Intensity
Theory fit
Experiment
1/d², 1/Å²
Spots list
Score = total scattered intensity × radial shape similarity

wafers.[14] Moreover, the protocol automates both multi-crystal data collection from a series of microcrystals contained in one, or more, sample holders and multi-position data collection from larger crystals. As we have previously shown,[13] Mesh&Collect is widely applicable, with data sets compiled using the method being used to solve the crystal structures of membrane proteins, to identify ligand binding sites in proteins and in structure determination using single-wavelength anomalous dispersion techniques.

3.2 Dozor

To recognize the presence of a diffraction pattern from a macromolecular crystal on a low dose mesh scan image, the program DOZOR is used.

The DOZOR algorithm (see Figure 3.3) is based on the assumptions that, in the case of diffraction from crystals of biological macromolecules, the intensity of Bragg spots is relatively weak and that most of the image pixels record only background scattering. Taking this into account, the mean values of the background intensity and its variations with scattering angle can be determined by the iterative summation of corresponding pixel intensities and the sequential rejection of outliers. The mean intensity of Bragg spots as a function of resolution is then defined as the difference between total intensity in a pixel and the corresponding background intensity. The average Bragg spot intensity determined in this way is a measure of diffraction strength and can be used for scoring/ranking of individual diffraction images. To improve the quality of the estimation of diffraction spot intensities, DOZOR approximates the experimental data by applying an isotropic Debye–Waller factor to the standard protein Wilson plot model.[4] The correspondence between experimental data and the unique pattern of average squared structure factor magnitudes is evaluated *via* the correlation coefficient and finally, a score of diffraction strength is estimated as the total averaged diffraction intensity multiplied by this coefficient.

3.3 Hierarchical Cluster Analysis (HCA)

Perhaps the most crucial aspect of multi-crystal data collection protocols such as Mesh&Collect is the best choice of partial data sets to combine in order to produce the final data set for structure solution and/or refinement.

Figure 3.3 The DOZOR analysis of single diffraction images starts with the determination of the distribution of background intensity (top left). This is used to produce a one-dimensional background function (top right) subsequently used to help identify the positions and intensities of any Bragg spots on the diffraction image (bottom left). The total scattered intensity in a diffraction image should follow the shape of a Wilson plot. The 'DOZOR score' for the diffraction image (bottom right) is estimated by convoluting the total scattered intensity in a diffraction image with the similarity of its distribution, as a function of resolution, to that of a Wilson plot.

Hierarchical cluster analysis (HCA) is a method aimed at finding homogeneous groups of data from a larger ensemble, based on some metrics of their similarity. This method has applications in all the branches of science, and is finding increasing use in MX, particularly in multi-crystal data collection experiments. However, there is some divergence of opinion as to the best metric to use in the HCA of partial data sets, with metrics based on either similarity of unit cell dimensions determined from partial data sets[19] or the correlation of diffraction intensities common to different partial data sets[20] being employed. For the Mesh&Collect procedure the latter approach is currently used.

Briefly, the HCA procedure used here relies on the distance between two data sets which is defined as $d(i,j) = \sqrt{1 - \mathrm{cc}^2_{(i,j)}}$, where $\mathrm{cc}^2_{(i,j)}$ is the correlation coefficient between the intensities of common reflections in data sets i and j. Once all distances between all possible data sets are obtained, HCA then builds clusters using linkage methods. Based on the distance matrix obtained above, the distance between two clusters X and Y is calculated with the Average Linkage Method as follows:

$$D(X,Y) = \frac{1}{N_X + N_Y} \sum d(i,j); \quad i \in X, \quad j \in Y$$

This allows one to maintain the relationship between the clustering distance and $\mathrm{CC}_{(i,j)}$. The result of the clustering procedure is best presented in a two-dimensional dendrogram (*i.e.* Figure 3.2, bottom left) that displays individual data sets on its x-axis and the similarity between them on its y-axis. Here, the lower a node linking partial data sets together, the more similar the partial data sets are.

3.4 Mesh&Collect in Practice

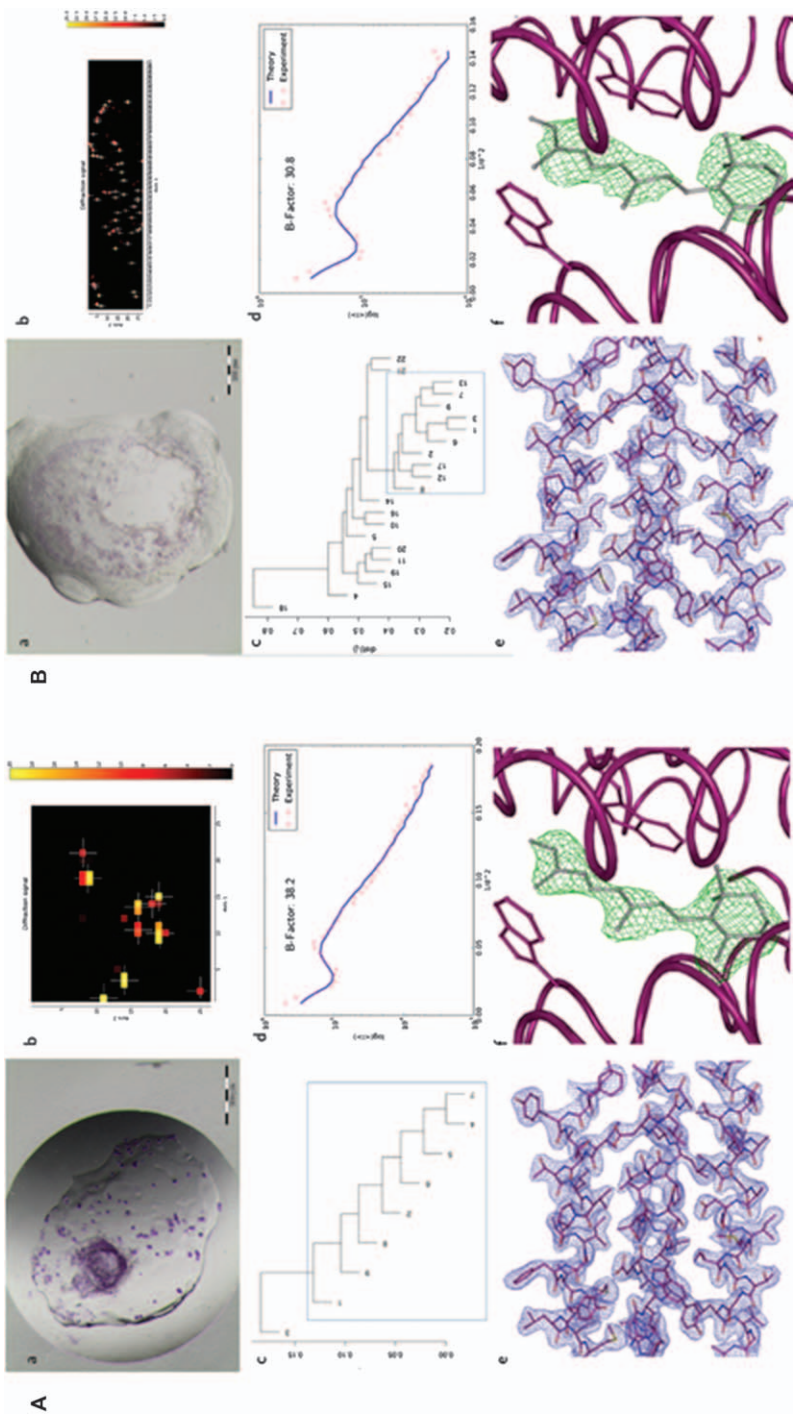
To prepare a cryo-temperature Mesh&Collect experiment, crystals are best mounted in a mesh-type sample holder (*i.e.* MiTeGen MicroMeshes; MiTeGen, USA), trying to avoid the stacking of crystals (Figure 3.2). Excess mother liquor should be removed and the sample holder flash-cooled. Once on the goniometer, the plane of the sample holder should be oriented such that it is perpendicular to the X-ray beam. At the ESRF, Mesh&Collect data collection is available *via* the MXCuBE beamline-control graphical user interface[21,22] from where it can be launched at the click of a button. Upon launching, the user defines the size of the X-ray beam to be used and the area (the 'mesh') over which the pre-interrogation of the sample holder should be performed (Figure 3.2). Diffraction images at each point in the mesh are then collected and analysed on-the-fly for protein diffraction using the program DOZOR (§3.2, Figure 3.3). Once the mesh scan has finished, the user receives a heat map (Figure 3.2), also stored in the ISPyB experiment-tracking system,[23] showing the points at which diffraction has been observed and colour-coded to indicate diffraction strength. Based on the heat map, the user then

defines, by setting a threshold for diffraction strength, how many partial data sets should be collected ($\pm 5^\circ$ total rotation range, 100 images per partial data set). These are then automatically collected and data processing automatically launched. Those data sets which are successfully processed are then scaled together and the $CC_{(i,j)}$ values are used to produce a dendrogram (Figure 3.2) using the HCA protocol outlined in §3.3 to determine which partial data sets should be combined to produce the data set used subsequently for structure solution and refinement (Figure 3.2).

3.4.1 Solving the Crystal Structures of Membrane Proteins with Very Small Crystals

Of the many stumbling blocks in the process leading to the solution of the crystal structures of membrane proteins, two are that the crystals obtained are often rather small and that crystallisation is in mesophases.[25,26] This means that, once mounted on sample holders at cryo-temperatures, the medium surrounding the crystal(s) is often opaque and the location, using visible light, of crystals is difficult. Mesh&Collect provides solutions to both these problems allowing the location of crystals *via* a preliminary X-ray mesh scan and the multi-crystal data collection and processing that is often required for the compilation of a complete data set for structure solution. In order to demonstrate the potential of Mesh&Collect for multi-crystal data collection from membrane protein crystals grown in a mesophase, we chose the example of bacteriorhodopsin (BR). Here, crystals were produced[27] and two separate batches used for Mesh&Collect experiments. For the first batch (BR1, Figure 3.4A), the crystals produced were rather large, with dimensions of $\sim 20 \times 20 \times 5$ μm^3. Here, the initial mesh scan (X-ray beam 20 μm in diameter; 3×10^{11} photons s^{-1}) located 10 well-diffracting crystals from which partial data sets were then collected, 9 of which were successfully processed automatically. After HCA, 8 of these were scaled and merged to produce the final data set (space group $P6_3$; $d_{min} = 2.3$ Å; completeness 92.7%; multiplicity 4.8; $CC_{(1/2)}$ 0.998; $\langle I/\sigma(I)\rangle$13.0; R_{pim} 0.05). Structure solution was carried out using Molecular Replacement (MOLREP[28]) using a previously determined crystal structure of BR[29] as a search model. Structure refinement was then carried out using REFMAC5[30] with alternating rounds of manual rebuilding in COOT,[31] resulting in $R_{cryst} = 0.232$ and $R_{work} = 0.239$. As can be seen in Figure 3.4, retinal co-factor molecules not included in the MR search model or in the initial stages of structure refinement were easily identifiable in 'omit' difference density ($mF_{obs} - DF_{calc}$; α_{calc}) maps.

The second batch of BR crystals (BR2) were much smaller, $\sim 5 \times 5 \times 2$ μm^3 in dimensions (Figure 3.4B). Pre-interrogation of the sample holder was performed using a Gaussian X-ray beam 10 μm in diameter, containing 1.5×10^{11} photons s^{-1}. The resulting heat map revealed the 59 positions from which partial data sets were collected. 38 of these could be


A
B
a
b
c
d
e
f
Theory
Experiment
B-Factor: 38.2
B-Factor: 30.8

automatically processed and, based on HCA, 10 were merged to construct the final data set (space group $P6_3$; $d_{min} = 2.54$ Å; completeness 96.7%; multiplicity 5.1; $CC_{(1/2)}$ 0.938; $\langle I/\sigma(I)\rangle$ 4.8; R_{pim} 0.18). Structure solution was carried out as for BR1, resulting in a final model yielding $R_{cryst} = 0.193$ and $R_{work} = 0.218$. As for BR1, retinal co-factor molecules not included in the MR search model or in the initial stages of structure refinement were easily identifiable in 'omit' difference density maps. This Mesh&Collect experiment thus clearly demonstrates its potential for use in constructing multi-crystal data sets from *very* small crystals.

The BR1 and BR2 experiments highlight another crucial requirement of automatic multi-crystal data collection and processing pipelines. That is, for polar space groups (*e.g.* $P6_3$ for BR) the first successfully processed partial data should then be used as a reference data set for the processing of the remaining partial data sets. This avoids introducing artificial merohedral twinning[32] into the merged data set and means that any twinning subsequently detected[33] is likely to be real. Determination of the true space group is thus made easier and calculated twinning fractions (0.06 in the case of BR1, 0.39 for BR2) can then be used in structure solution and refinement processes.

3.4.2 Multi-crystal Data Collection for Ligand Binding Studies

The examples of BR1 and BR2, where omit difference density for ligands not included in MR search models or during initial refinement of the crystals structures are clearly visible and assignable, suggest that multi-crystal data collection pipelines such as Mesh&Collect can also play a significant role in fragment screening projects carried as an aid to structure-based drug design.[34,35] Usually, fragment screening campaigns rely on the production of relatively large, robust, single crystals with introduction of ligands either *via* co-crystallisation or soaking experiments.[36] However, it has been shown that in the latter, smaller crystals require reduced fragment/ligand soaking times

Figure 3.4 Mesh&Collect for multi-crystal data collection from membrane protein crystals grown in a mesophase. We chose the example of bacteriorhodopsin (BR). A: (BR1) the initial mesh scan of a sample holder containing several crystals of bacteriorhodopsin $\sim 20\times20\times5$ μm^3 in dimensions located 10 well-diffracting crystals from which partial data sets were collected. After HCA, 8 of these were scaled and merged to produce the final data set used in structure determination. B: (BR2) Microcrystals $\sim 5\times5\times2$ μm^3 in size of bacteriorhodopsin were mounted in the same sample holder. The Mesh&Collect heat map revealed the 59 positions from which partial data sets were then collected. 38 of these were automatically processed and 10 merged to construct the final data set used in structure solution/refinement.
Both figures from Zander *et al.*, 2015[13] and reproduced with permission of the International Union of Crystallography (http://journals.iucr.org/).

to obtain the same occupancy of a fragment/ligand[37] and thus multi-crystal data collection based around the production of micro-crystals might prove to be a viable and efficient alternative to the traditional approach of producing relatively large single crystals for fragment screening/ligand binding studies. To test this idea, we carried out a Mesh&Collect experiment (Figure 3.5) using relatively small crystals of thaumatin crystallised in the presence of tartrate (see ref. 13 for details). In our experiment, an entire crystallisation drop containing many crystals (dimensions $\sim 40 \times 40 \times 60\ \mu m^3$) was harvested on a micromesh loop and mounted at 100 K on ESRF beamline ID29.[38] The Mesh&Collect workflow was launched using an X-ray beam 10 μm in diameter containing 8.7×10^{11} photons s^{-1}. Partial data sets were collected from 100 well-diffracting positions identified in the initial mesh scan. 78 of the partial data sets could be automatically integrated and, after HCA, 74 combined to build the final data set (space group $P4_12_12$; $d_{min} = 1.2$ Å; completeness 99.4%; multiplicity 51.3; $CC_{(1/2)}$ 0.999; $\langle I/\sigma(I)\rangle$ 48.7; R_{pim} 0.034).

Structure solution was carried out using MR with PDB code 4AXU,[39] stripped of water molecules and ligands, as the search model. Structure refinement, carried out using REFMAC5 alternated with rounds of manual rebuilding in COOT, produced a final model yielding $R_{cryst} = 0.133$, $R_{work} = 0.151$. As can be seen in Figure 3.5, a tartrate molecule bound to thaumatin is very visible and easily identifiable in an 'omit' difference density ($mF_{obs} - DF_{calc}$; α_{calc}) map calculated at the end of the refinement procedure.

3.4.3 *De novo* Structure Solution using Mesh&Collect

To investigate the possibilities of the Mesh&Collect workflow to produce data suitable for *de novo* structure solution techniques that exploit anomalous scattering, Zander *et al.*[13] presented results on two different systems: crystals of thermolysin from *Bacillus thermoproteolyticus* and crystals of the selenomethionyl derivative of the MAEL domain from *B. mori maelstrom.* Both crystal structures were phased successfully using anomalous signals from Zn^{2+} and Se respectively. The former experiment demonstrated that Mesh&Collect can produce diffraction data of sufficient quality to exploit relatively small anomalous signals ($\langle \Delta F/F \rangle \sim 2\%$) in SAD structure solution, the latter demonstrated that Mesh&Collect can be extended to samples contained in more than one sample holder. Below we show that Mesh&Collect workflow can also be successfully used in S-SAD,[40,41] I-SAD[42,43] and I-SIRAS experiments.[43]

3.4.3.1 The Solution of the Crystal Structure of Thaumatin using S-SAD

S-SAD exploits the anomalous signal of sulfur atoms present in the amino acid sequences (cysteine or methionine) of most protein molecules. However, in S-SAD experiments carried out on 'standard' synchrotron-based MX

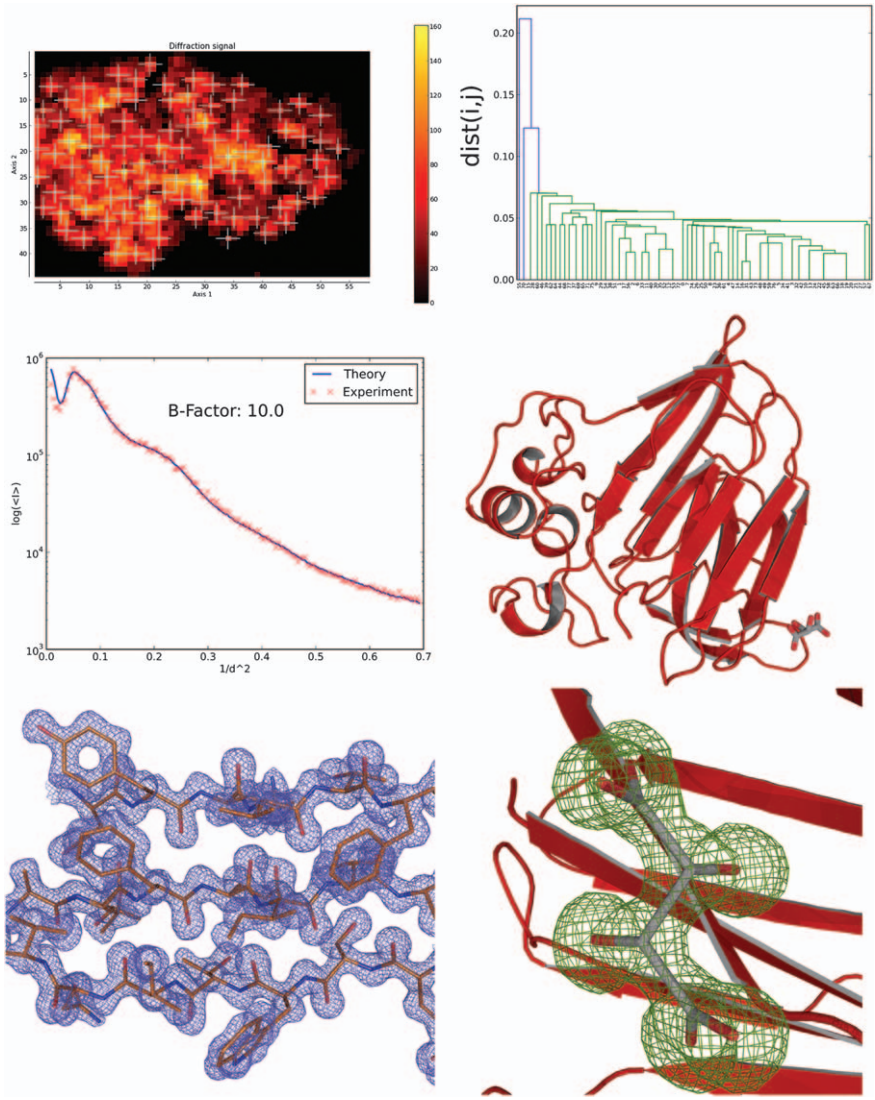


Figure 3.5 Mesh&Collect for ligand binding studies. Dozor analysis of the sample holder was used to guide the collection of 100 partial data sets of which 78 could be automatically processed. HCA was then used to guide the choice of the merging of 74 partial data sets to produce the final data set for structure solution/refinement. An 'omit' difference density map calculated at the end of the refinement procedure clearly shows the presence of a tartrate molecule bound to the protein.
Adapted from Zander *et al.*, 2015[13] and reproduced with permission of the International Union of Crystallography (http://journals.iucr.org/).

beamlines, the anomalous signals ($\langle \Delta F/F \rangle$) that can be induced are rather small, often in the order of 1–2%, and to measure such small signals requires very high data quality.[44,45]

For our Mesh&Collect experiment (Figures 3.6 and 3.7, Table 3.1) aimed at the S-SAD solution of the crystal structure of thaumatin, many crystals (maximum dimension between 10 μm and 30 μm) were loaded onto a sample holder, flash-cooled and mounted onto a goniometer. In the initial mesh scan of the sample holder, 95 well-diffracting positions were identified from each,

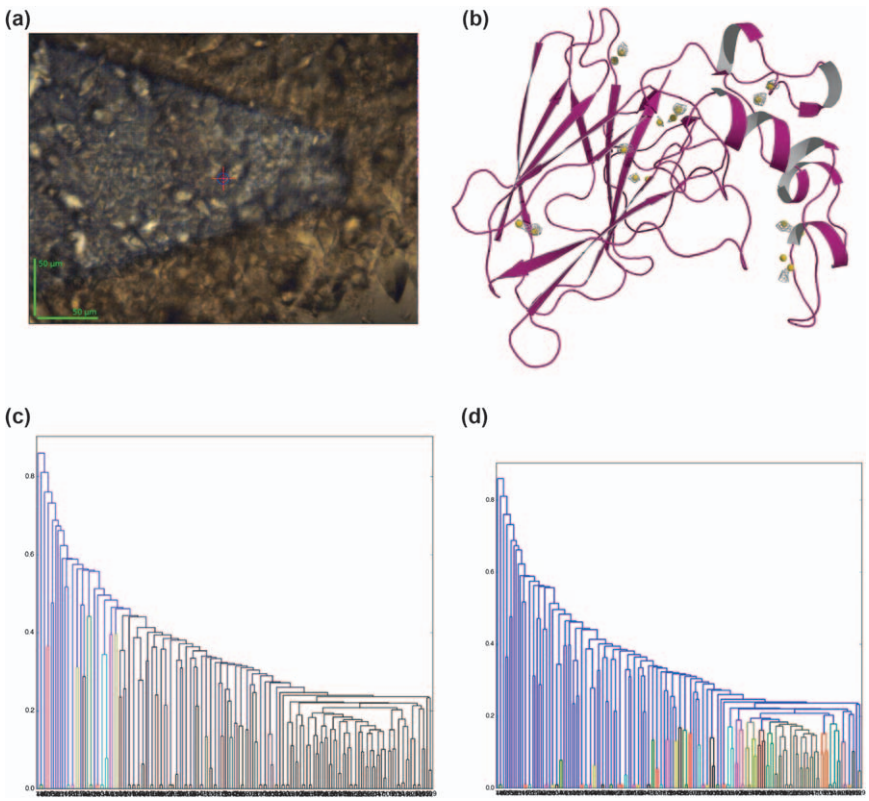


Figure 3.6 A Mesh&Collect experiment aimed at the S-SAD structure solution of thaumatin using the inverse beam method; (a) close-up of the thaumatin crystal slurry mounted on a mesh support; (b) the refined model of thaumatin obtained from the experiment (the main chain of the molecule is shown in secondary structure representation (magenta), the sulfur atoms as yellow spheres. The peaks in the anomalous difference maps (ΔF_{ano}, α_{calc} +90°; contoured at 3×r.m.s.) corresponding to the positions of the S atom are shown in grey chicken wire); (c) HCA dendrogram showing the cluster of partial data sets (139 in total) merged to produce a data set resulting in successful structure solution (see Figure 3.7); (d) HCA dendrogram showing a cluster of partial data sets (26 in total) which, when merged, produced a data set that did not allow successful structure solution (see §3.5).

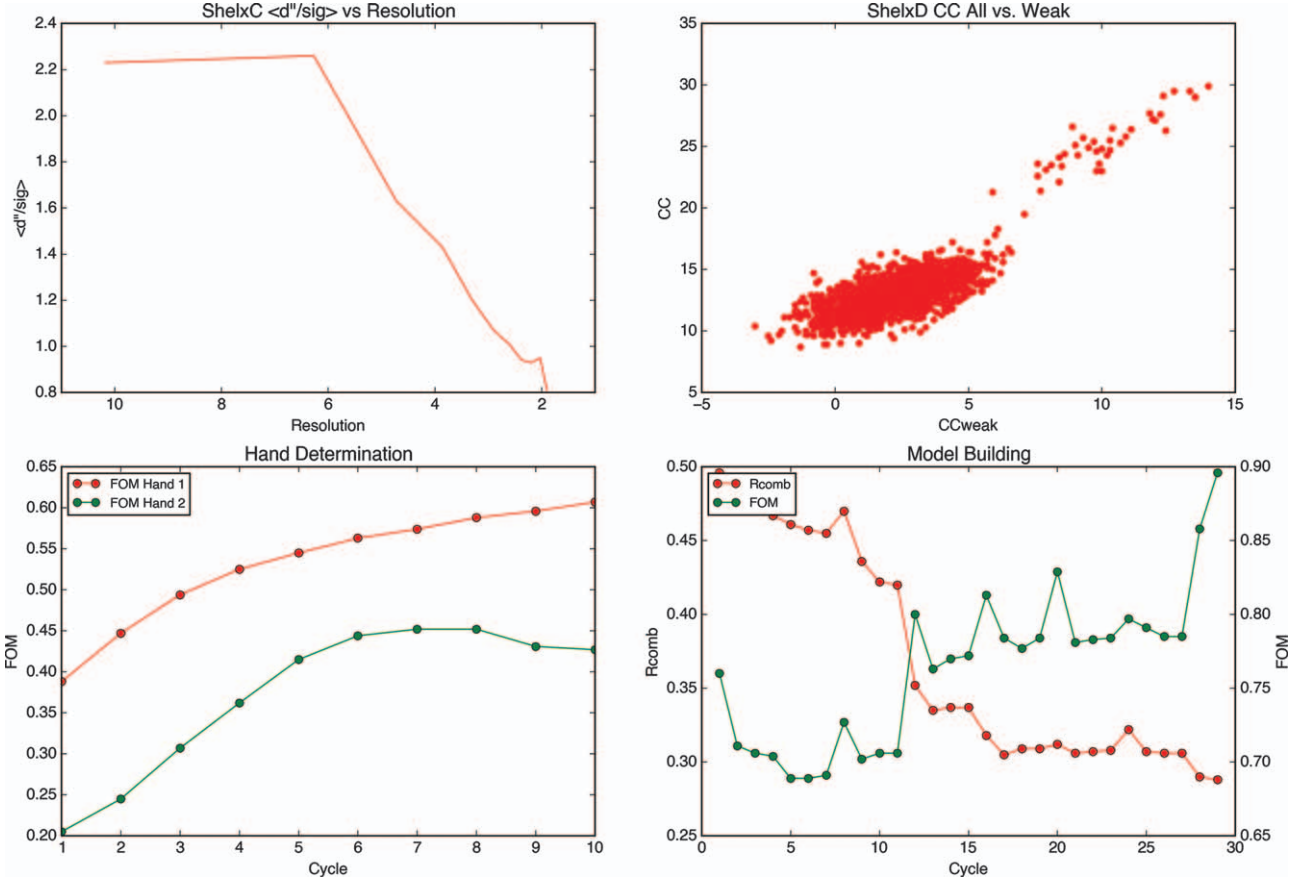


Figure 3.7 Successful S-SAD structure solution using data collected with Mesh&Collect. From top left: A plot of $\langle \Delta F_{\rm ano}/\sigma(\Delta F_{\rm ano})\rangle$ *vs.* resolution for the final data set obtained; successful S atom substructure is indicated by the bimodal distribution of the plot of $CC_{\rm all}$ *vs.* $CC_{\rm weak}$ for 2000 trials; the S atom substructure obtained clearly allows discrimination of the correct hand to use in phase calculations; plots of FOM and $R_{\rm comb}$ as a function of model building cycle in CRANK2.

Table 3.1 Statistics for two data sets, produced using Mesh&Collect, intended for use in the S-SAD phasing of the crystal structure of thaumatin. See main text (Sections 3.4.3 and 3.5) for discussion.

Wavelength (Å)	2.0	
Space group	$P4_12_12$	
Number of merged partial data sets	139	26
Unit cell: $a = b$ (Å), c (Å)	57.87, 150.34	57.83, 150.30
Resolution range (Å)	19.74–1.90 (1.94–1.90)	19.73–2.00 (2.05–2.00)
Total no. of reflections	1 391 160	251 087
No. of unique reflections	21 026	18 029
Completeness (%)	99.9 (99.8)	99.7 (98.0)
Multiplicity	66.2 (12.4)	13.9 (4.6)
B factor, Wilson plot (Å^2)	21.7	17.5
$CC_{1/2}$	0.998 (0.599)	0.996 (0.712)
$\langle I/\sigma(I)\rangle$	23.3 (0.7)	14.6 (2.2)
R_{pim}	0.032 (1.004)	0.039 (0.381)
R_{meas}	0.262 (3.609)	0.150 (0.872)

of which 10° of data (5° forward sweep, 5° reverse sweep) were automatically collected using an inverse beam protocol.[46,47] Of the 190 partial data sets collected, 183 were automatically integrated and a cluster of 139 selected using HCA (cluster distance ~0.42; $\langle CC(i,j)\rangle \sim 0.9$) for the compilation of the final data set (Table 3.1). Using this data set, S atom substructure determination, phasing, density modification and model building was successfully performed with the CRANK2 automated structure solution pipeline.[48]

3.4.4 Mesh&Collect for the I-SAD/I-SIRAS Solutions of the Crystal Structure of the KR2 Light-Driven Sodium Pump

The Mesh&Collect experiments described above for BR (§3.1) clearly demonstrate the potential of Mesh&Collect for construction, from (very) small crystals, of a multi-crystal data set of sufficient quality to allow the MR-based solution of the crystal structures of membrane proteins. However, if no model of a previously-determined crystal structure of either the protein itself or of a structural homologue is available, as for all problems in MX, *de novo* phasing of the crystal structure must be carried out. Recently, it has been suggested that iodide derivatisation of membrane protein crystals might prove a universal method for their *de novo* structure solution using either I-SAD (data collection from derivative crystals only) or I-SIRAS (data collection from both native and derivative crystals).[49]

To investigate whether Mesh&Collect protocols might be suitable for such experiments, we carried out Mesh&Collect-based I-SAD and I-SIRAS experiments (Figure 3.8) using crystals, grown in a mesophase, of the light-driven sodium pump KR2.[50,51] In all cases crystals had maximum dimensions of 10–15 μm. For the I-SAD experiment, data were collected at $\lambda = 1.85$ Å and crystals were harvested on a micromesh loop, flash-frozen and mounted on

the goniometer of ESRF beamline ID23-1.[52] The initial mesh scan revealed 99 well-diffracting positions from each, of which 20° of data were collected using the inverse beam methodology[46,47] (10° forward sweep, 10° reverse sweep) resulting in 198 partial data sets. 171 of these were successfully processed and, following HCA, 136 were merged to produce the final data set (see Table 3.2) that could be used in a very straightforward I-SAD structure solution (Figure 3.8d–g). In order to apply I-SIRAS as a structure solution method, only 39 partial data sets were merged and used in structure solution (Table 3.2 and Figure 3.8h–k) in combination with a native data set collected in a standard fashion (*i.e.* from a single crystal).

3.4.5 Mesh&Collect at Room Temperature

The major experimental requirement for Mesh&Collect experiments is that multiple crystals are fixed on some sort of solid support. This allows both the determination of the positions of crystals during the initial mesh scan and the ability to return to these positions for subsequent collection of partial data sets. While this is very straightforward when samples are cryo-cooled, it is no longer the case at room temperature, especially if the crystals are mounted on a standard nylon loop or micromesh. Here, multiple crystals floating in mother liquor will tend to cluster together through the action of gravity and, even if this does not happen, rotation of the sample holder during the data collection steps might cause crystals to move from their initially located positions. Fortunately, however, crystals sandwiched between Si_3N_4 membranes,[18] COC wafers[14] or other similar supports have been shown to be sufficiently reliably fixed in place for serial MX experiments. In particular, both provide mechanical stability and prevent rapid dehydration of samples while, at the same time, producing low background scattering. We therefore tested the potential of using crystals mounted in a Si_3N_4 membrane sandwich in Mesh&Collect experiments (Figure 3.9). Here, many crystals (~10 μm in maximum dimension) of tetragonal hen egg-white lysozyme[18] were deposited on the sample support and Mesh&Collect carried out on the ESRF beamline ID30B.[54] The initial mesh scan (10 μm beam size) revealed 100 well-diffracting positions at each, of which partial data sets comprising 5° total oscillation range were subsequently collected. 98 of these were successfully processed and, after HCA, a cluster of 36 data sets were merged to produce the final data set (space group $P4_12_12$; $d_{min} = 2.0$ Å; completeness 99.0%; multiplicity 12.6; $CC_{(1/2)}$ 0.99; $\langle I/\sigma(I)\rangle$7.7; R_{pim} 0.18). MR structure solution and refinement resulted in a final model yielding $R_{cryst} = 0.222$ and $R_{free} = 0.275$.

3.5 The Pitfalls of HCA

Once a sample holder is mounted on the goniometer, all steps of the Mesh&Collect workflow up to, and including, the processing of the partial data sets collected are carried out in a semi-automatic fashion. However, the

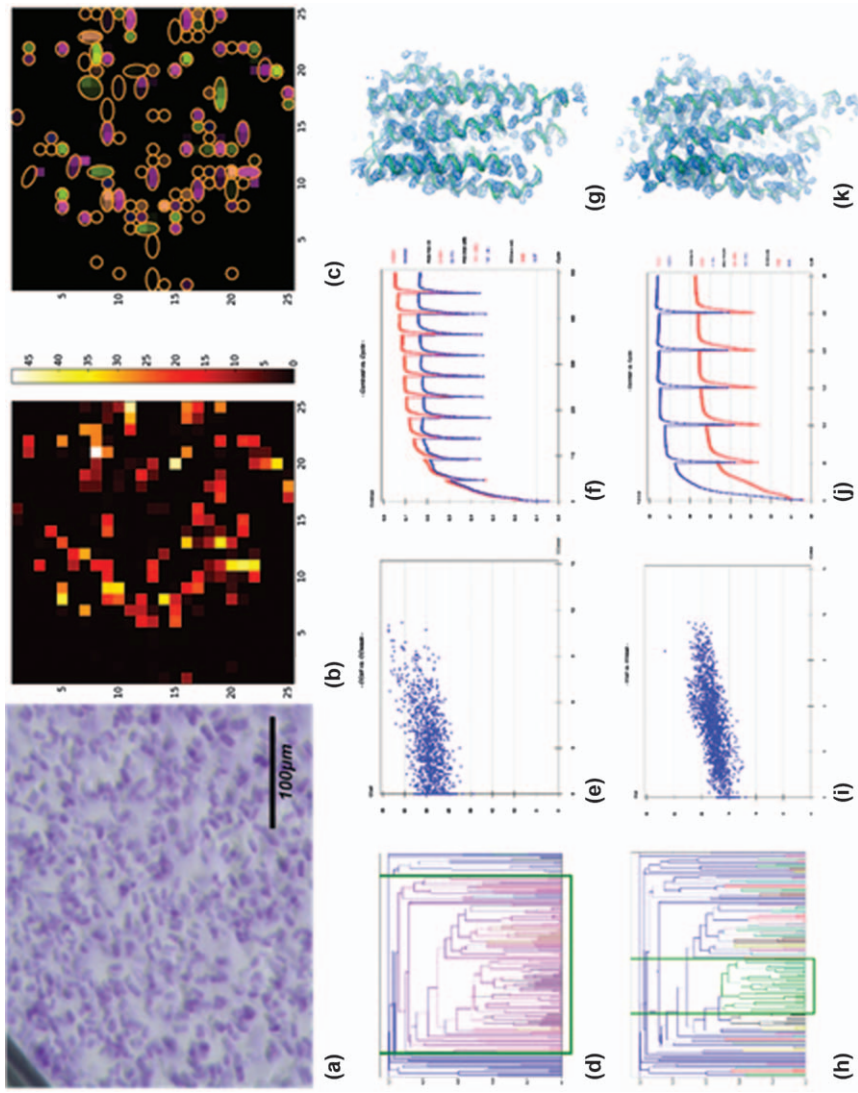

100μm
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)

Table 3.2 Data quality statistics for the Mesh&Collect multi-crystal data sets used in successful I-SAD and I-SIRAS phasing experiments to determine the crystal structure of the light-driven sodium pump KR2.

Method	I-SAD	I-SIRAS
Wavelength (Å)	1.85	
Space group	I222	
Number of merged partial data sets	136	39
Unit cell: a (Å), b (Å), c (Å)	40.6, 83.6, 233.5	41.0, 84.0, 234.3
Resolution range (Å)	60–2.8 (3.0–2.8)	60–2.9 (3.1–2.9)
Total no. of reflections	469 213 (85461)	124 613 (22610)
No. of unique reflections	18 928 (3542)	17 331 (3171)
Completeness (%)	100 (100)	99.8 (100)
Multiplicity	24.79 (24.13)	7.19 (7.13)
$CC_{1/2}$	0.998 (0.701)	0.993 (0.562)
$\langle I/\sigma(I)\rangle$	10.64 (2.27)	7.71 (1.69)
R_{merge}	0.617 (2.207)	0.238 (1.546)
R_{meas}	0.630 (2.254)	0.256 (1.666)
SigAno	1.21 (0.78)	1.03 (0.73)
CC_{anom}	0.39 (0.00)	0.29 (−0.01)

choice of partial data sets to merge to produce the final data set for downstream analysis is not yet straightforward enough to automate. Unfortunately, however, this aspect is the most crucial, and failure to choose the correct partial data sets to merge can have a huge bearing on the outcome of an experiment.

An excellent illustration where this is the case is a Mesh&Collect experiment using crystals of monoclinic lysozyme (ref. 13; Figure 3.10). Here, initial pre-interrogation of the sample holder revealed 54 positions at which partial data sets were automatically collected and of which 40 were subsequently processed, scaled and analysed by HCA. Merging 39 of these partial data sets resulted in a final data set with 95% overall completeness to $d_{min} = 1.6$ Å with $R_{pim} \sim 16\%$ over the entire resolution range. However, at a value of ~0.9, the HCA distance value for the cluster of partial data sets

Figure 3.8 The Mesh&Collect-based I-SAD/SIRAS phasing of the crystal structure of the KR2 light-driven sodium pump. (a) The crystals used in the experiment were of maximum dimension 10–15 μm; (b) The heat map generated following the initial mesh scan of the sample holder; (c) The MeshBest analysis (see §3.8) of the mesh-scan with the size and position of individual crystals. Different crystals/different zones of homogeneous diffraction that are bordering with each other are illustrated as ellipses or circles; (d)–(g) I-SAD phasing: the HCA dendrogram of the partial data sets collected (cluster chosen for merging indicated by green box); heavy atom substructure determination and phase calculation and extension using ShelxC/D/E;[53] the electron density map, contoured at 2.0×r.m.s. level, superimposed with the initial model obtained; (h)–(k) I-SIRAS phasing: the HCA dendrogram with the cluster chosen for I-SIRAS phasing indicated by a green box; heavy atom substructure determination, phase calculation and extension; the resulting electron density map, contoured at 2.0×r.m.s. level, superimposed with the initial model obtained.

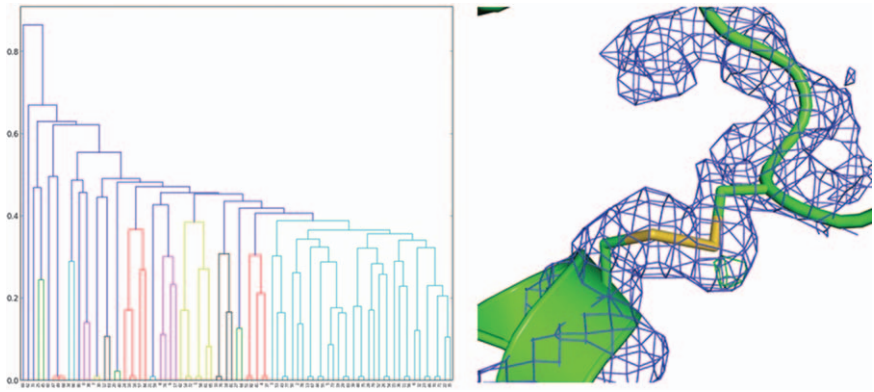


Figure 3.9 Left: The HCA dendrogram obtained from a room temperature Mesh&Collect experiment using very small crystals of tetragonal hen egg-white lysozyme. The cluster of 36 data sets merged to produce the final data set is highlighted in cyan. Right: A detail of the electron density map (blue) obtained after structure solution and refinement. The region shown is the Cys75-Cys95 disulfide bridge where difference electron density (green) suggests little sign of specific radiation damage.

chosen for merging is relatively high, corresponding to an average correlation coefficient for common intensities ($\langle CC_{(i,j)} \rangle$) between partial data sets of ~0.4 and suggesting considerable non-isomorphism between at least some of the partial data sets collected. We therefore compiled a second data set based 21 partial data sets with cluster distance of ~0.4, corresponding to $\langle CC_{(i,j)} \rangle \sim 0.9$. Here, due to the relatively low symmetry of the crystals, overall completeness to $d_{min} = 1.6$ Å was only 85%. However, the overall quality of the compiled data set was significantly improved (R_{pim}(overall) ~8%). Using either of the two data sets, the crystal structure could be straightforwardly solved using MR. During refinement using the data set based on a cluster distance of ~0.4, examination of difference electron density ($mF_{obs} - DF_{calc}$, α_{calc}) maps after partial model refinement revealed clear difference density which could be unequivocally assigned as a nitrate ion binding to the protein (Figure 3.10e). However, this was not the case in difference density maps calculated when refining the crystal structure using the final data set based on a cluster distance of ~0.9. Here, this feature is present but is ''blurred out'' and would not have been assigned as a nitrate ion in the final refined model (Figure 3.10f).

A second example of the importance of choosing the best partial data sets to use in the compilation of a final data set is from the Mesh&Collect experiment, described in §3.4.3.1, aimed at the *de novo* solution, using the S-SAD technique, of the crystal structure of thaumatin. Here, the merging of 139 data sets selected using a HCA cluster distance of ~0.4 ($\langle CC_{(i,j)} \rangle \sim 0.9$) (Figure 3.6) produced a final data set (Table 3.1) which allowed successful structure solution (Figure 3.7). However, as the example of monoclinic lysozyme outlined above suggested that better results might be

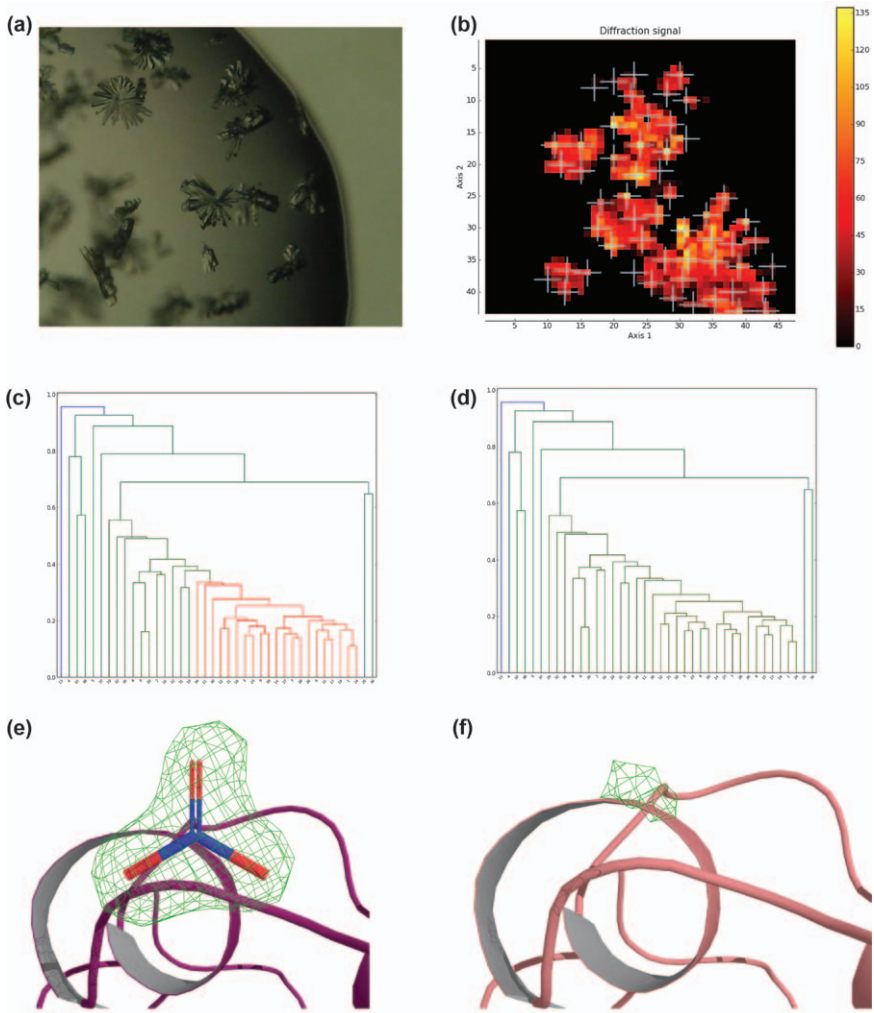


Figure 3.10 Different results obtained from a Mesh&Collect experiment with monoclinic lysozyme crystals: (a) The crystals used in the experiment; (b) The DOZOR heat map of the initial mesh scan; (c) The HCA dendrogram with partial data sets clustering at a distance of ∼0.4 highlighted in red; (d) Dendrogram generated by hierarchical cluster analysis with the partial data sets clustering at a distance ∼0.9 highlighted in green; (e) Omit difference density, contoured at the 3×r.m.s. level, after structure solution (MR) and refinement using the low distance cluster data set; (f) Omit difference density, contoured at the 3×r.m.s. level, after structure solution (MR) and refinement using the high distance cluster data set.

achievable using a cluster of more isomorphous data sets with a higher $\langle CC_{(i,j)} \rangle$ so a second data set was compiled based on a clustering distance of ∼0.18 ($CC_{(i,j)} > \sim 0.98$) (Figure 3.6(d), Table 3.1). Here, however, all attempts

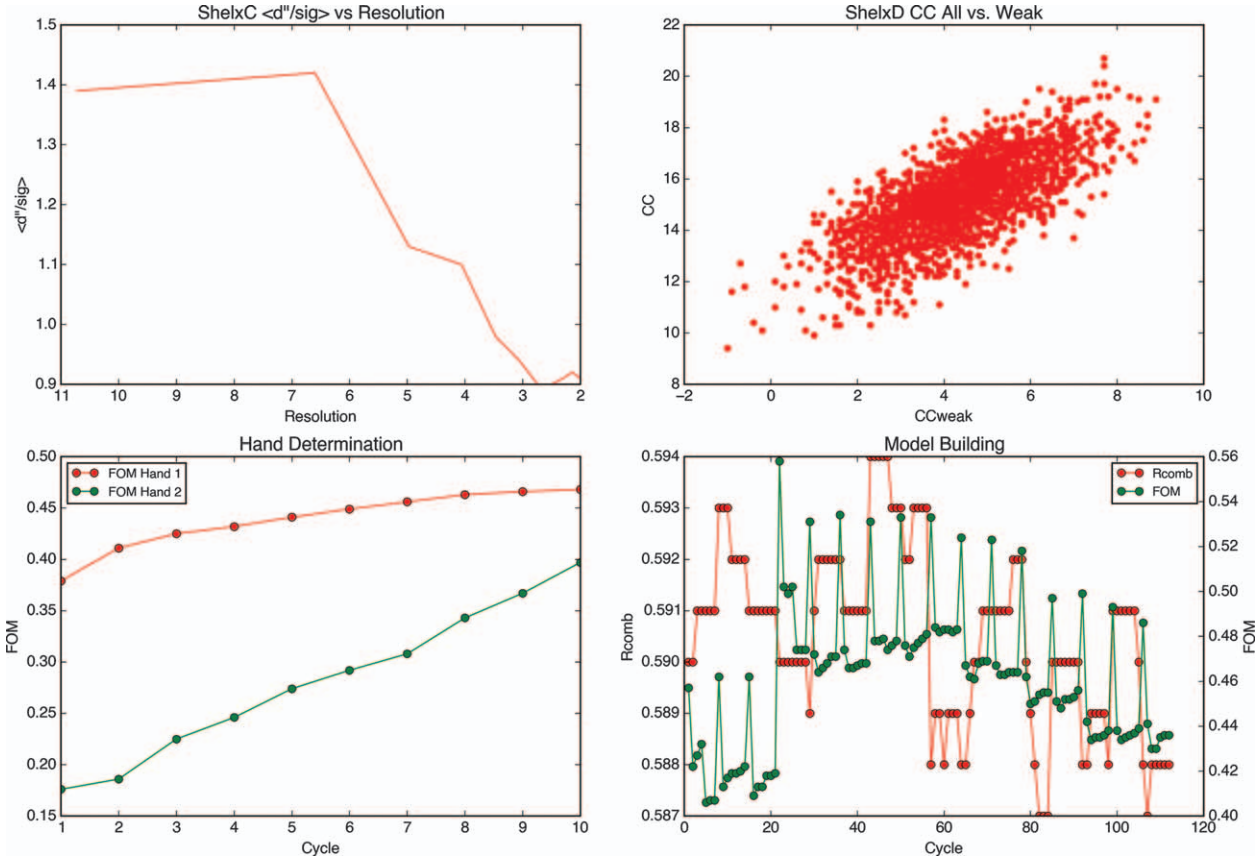


Figure 3.11 Attempted multi-crystal S-SAD phasing using the small cluster of partial data sets (26 in total) identified using HCA (see Figure 3.6). The lower values, particularly in the low-resolution bins, of $\langle \Delta F_{\rm ano}/\sigma(\Delta F_{\rm ano})\rangle$ in the resulting data set clearly do not allow successful structure solution. See Figure 3.7 for a comparison.

at structure solution were unsuccessful (Figure 3.11), probably because $\langle \Delta F_{ano}/\sigma(\Delta F_{ano}) \rangle$ at low resolution was much lower than for the initially compiled data set.

3.6 ccCluster

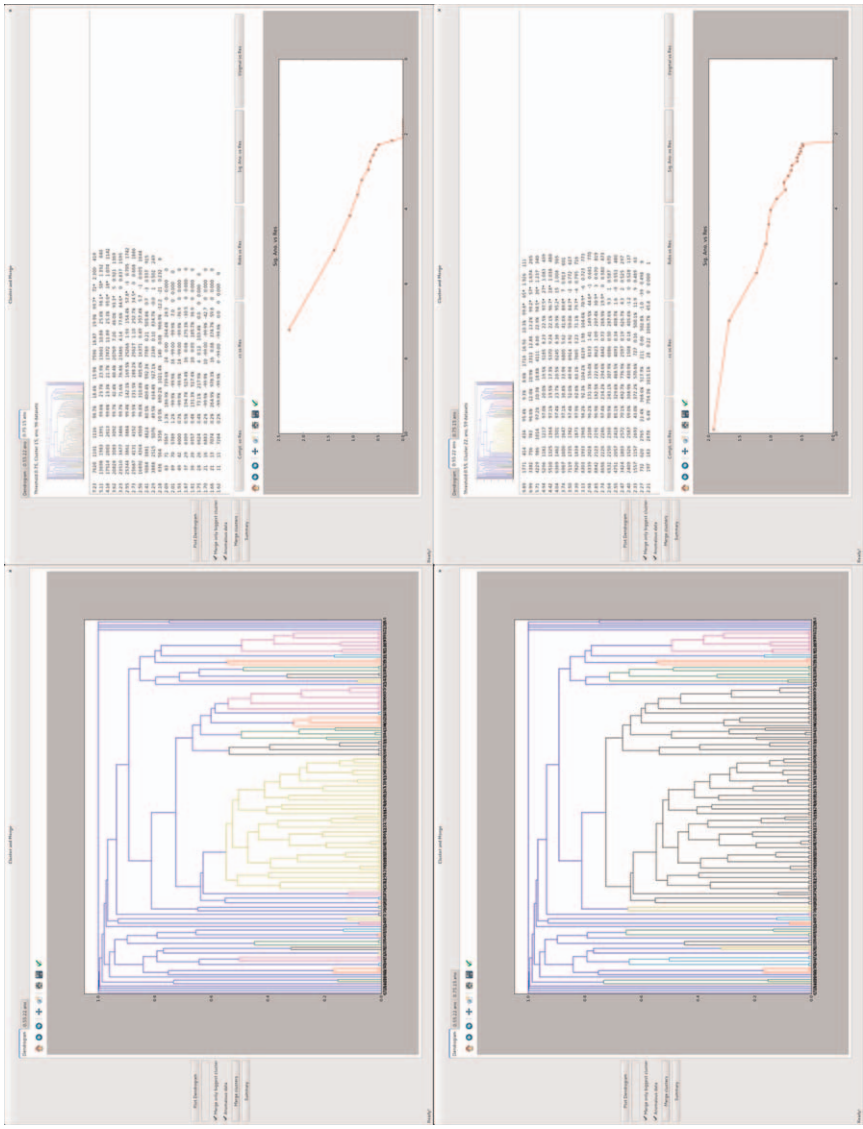
The examples briefly presented in §3.5 clearly show that the optimum choice of partial data sets to merge in a Mesh&Collect experiment – or any other multi-crystal MX experiment, for that matter – is, even with the help of techniques such as HCA, not straightforward and will often come down to a trial and error process in which several combinations are tried before the best compiled data set is arrived at. However, and somewhat unfortunately, if the first cluster chosen for merging does not result in a successful conclusion to an experiment, the combination of different data sets usually requires manual intervention in the form of the generation of new input files and the examination of the resulting log files to see if a particular aspect of data quality has improved.

To streamline this rather labour-intensive task we have developed ccCluster (https://github.com/gsantoni/ccCluster), a python-based program for HCA, which provides automated data merging capabilities with a graphical user interface for more interactive operation (Figure 3.12).

The novelty with respect to the method of Giordano *et al.*[20] is that calculation of $CC_{(i,j)}$ does not rely on running XSCALE but is performed directly using cctbx modules.[55] HCA is then performed with NumPy modules (http://www.numpy.org/). Once the data sets to merge are chosen from the dendrogram, XSCALE is used to merge them. Threshold determination (*i.e.* the linkage level for the cluster of partial data sets to be merged) is performed in 3 ways in ccCluster, with different levels of complexity: the program gives an initial estimate based on the number of data sets contained in the biggest cluster; users specify a different value from a command line option; users specify a different value by clicking on the dendrogram from the GUI. In all three cases, once the cluster(s) of partial data sets to be merged is chosen, input files for XSCALE are automatically created and run. Importantly, after XSCALE has been run, both the log file and relevant statistics are accessible from the ccCluster GUI (Figure 3.12), allowing users to take an informed decision as to whether data quality is good enough or whether a new threshold should be tried.

3.7 Merging of Partial Data Sets Using Genetic Algorithms

As illustrated in §3.5, using HCA to choose the best combination of Mesh&Collect partial data sets to merge to produce the final data set can often be a question of trial and error. Moreover, if the correct combination of partial data sets is not used, this can adversely affect the outcome of an experiment. However, in random-orientation multi-crystal experiments,



such as Mesh&Collect, which are based on the combination of relatively small partial data sets to produce the final data set, the results of HCA based on $\langle CC_{(i,j)} \rangle$ values calculated between partial data sets may be misleading due to the relatively few common reflections between some of the partial data sets. Thus, a more robust method is likely required for the optimal identification of partial data sets to be merged. One recent approach that has been developed at the ESRF is based on the idea that data quality statistics themselves can be optimized and can be used in the search for the global optimal grouping of partial data sets.[56] To this end, we have made use of a global optimization algorithm known as a genetic algorithm (GA, Figure 3.13) that is based loosely on the principles of biological evolution. The key steps in formulating the grouping of partial data sets in the context of a GA are: (1) the encoding of grouping into computational chromosomes; (2) the creation of a suitable fitness target based on the quality of the final data set. Once these steps have been performed, apart from tuning parameters, the optimization protocol itself can be implemented using a wide variety of off-the-shelf software components. In our case, we have used the *DEAP* package (https://github.com/deap/deap) which offers a complete set of tools for the straightforward development of a GA.

3.7.1 Grouping Partial Data Sets into Chromosomes

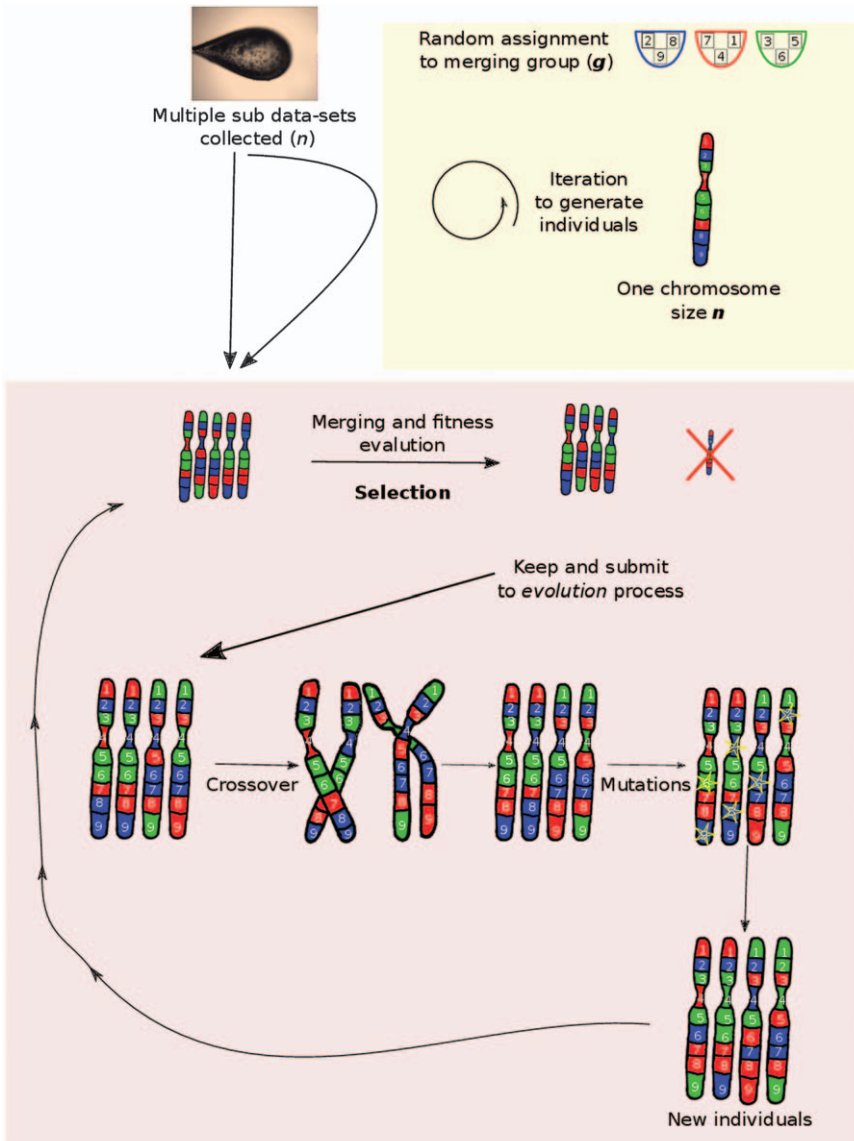
Each chromosome is constituted by a number n of genes, where n is the number of sub-data sets. Genes are randomly assigned to an integer corresponding to a merging group g (in our implementation, three merging groups $g = 3$ by default). The range of integers is limited to $1 \ldots g$. A chromosome is a simple way to encode the partial data sets into different sub-groups of merging and, critically, it implies that there is no overlap between the merging groups; any given partial data set cannot belong to multiple groups of merging in the same chromosome. This chromosome is then assigned to an individual, and multiple individuals exist in the population.

3.7.2 Fitness Evaluation

Our implementation of GA 'fitness' is based on final data set merging statistics. The first step consists of scaling together all the genes with the same group *g*. This is done using XSCALE.[24] The XSCALE log file (XSCALE.LP) is parsed for data-quality statistics and from these fitness, based on a combination of criteria including $\langle I/\sigma(I) \rangle$, R_{meas}, $CC_{1/2}$, completeness, multiplicity

Figure 3.12 Using ccCluster to rapidly merge partial data sets. Top: Initial HCA dendrogram from a Mesh&Collect experiment. ccCluster will automatically suggest the merging of the largest cluster (highlighted in yellow) and, when this is done, provide statistics (right panel) in the form of the XSCALE.LP log file and the graphical output of data quality statistics *vs.* resolution. Bottom: At the click of a mouse users can define new cluster distances for the merging of different combinations of partial data sets.

and CC_{anom}, is calculated. While default parameters usually produce high quality merged data sets, thanks to the flexibility of the GA it is possible to weight different parameters more strongly for specific downstream applications. For example, in anomalous phasing or molecular replacement one might give more weight to CC_{anom} or $\langle I/\sigma(I)\rangle$, respectively. Because the relative weights of the different terms can in some cases be difficult to balance (analogous to the case of finding appropriate weights for X-ray and geometry terms in the refinement of atomic models against X-ray data), we



have recently implemented a simple method to roughly equilibrate the different weights. Here the user sets a desired value for R_{meas} in the inner resolution shell and a short pre-optimisation is carried out using GA. The resulting $\langle I/\sigma(I)\rangle$, anomalous signal (SigAno in XSCALE) and $CC_{1/2}$ are then scaled to the R-term. Any user-requested up-weighting of these terms are then applied to the equilibrated terms. More sophisticated target functions are now under development that can be applied to multiple data sets to, for example, optimise isomorphous or RIP (Radiation damage Induced Phasing[57,58]) signals. This simply requires the addition of terms that describe the differences between two data sets to the scoring function (for example R_{iso}). This could be of particular interest if successful, as it opens up the possibility of single sample holder isomorphous replacement or RIP experiments, exploiting the inherent differences in heavy atom occupancy in soaking experiments.

3.7.3 GA Optimisation

Evolution is achieved by crossover between individuals and mutations (mutations resulting in modification of merging group affiliations). All GA is done with a simple pre-built algorithm with DEAP called EASimple, which applies uniform crossover ($p = 0.05$) and uniform mutation ($p = 0.05$) and a tournament of selections (tournament size = 3). Crossover and mutation probability can be specified by the user and the defaults are set to 0.3 and 0.6, respectively.

3.7.4 Case Study: LUX

In order to demonstrate the improvement of the merging statistics for a difficult test case, as well as the effects of changing different parameters, here we expand upon LUX, one of the examples discussed in Zander *et al.* (2016).[56] Merging all partial data sets results in a compiled data set with

Figure 3.13 Schematic example of the employment of GA in the context of merging SSX data for three merging groups. *n* Partial data sets are randomly divided into groups. The segregation of data sets is described in a chromosome (upper right). This is done multiple times to generate a population of chromosomes/individuals (each individual only has a single chromosome in our implementation). For each chromosome in every individual, three scaling runs in XSCALE are then performed, one for each group. The merging statistics are then converted to fitness scores, and the individual receives the fitness for the best scoring group (it is also possible to use the average fitness). In this case, individual 5 is removed from the population because of lower fitness and replaced with a new individual. The DEAP built-in mutation and crossover genetic modifiers are then applied followed by cycling back to the scoring step. The background colour indicates the source of the chromosome.

extremely high R_{meas} values, as well as rather low values of $CC_{1/2}$ and $\langle I/\sigma(I)\rangle$. Varying the weighting parameters after the initial pre-optimisation of weights produced merging data of much higher quality, and can be tuned to improve R_{meas}, $CC_{1/2}$ or $\langle I/\sigma(I)\rangle$, as shown in Table 3.3. The choice of which merging parameters to use is the topic of some debate, even taking into consideration recent work linking merging statistics to model quality.[59] In our experience, *R* values, even those including multiplicity weighting such as R_{meas}[60] or R_{pim},[61] are not as predictive of improvements to model quality compared to $CC_{1/2}$ or $\langle I/\sigma(I)\rangle$. However, it has become apparent that more tools for evaluating the utility of different merging statistics (such as paired refinements) are necessary to fully explore this subject.

3.8 MeshBest

Mesh&Collect is a user-friendly pipeline for the rapid collection of partial data sets from many crystals contained on the same or different sample holder(s) and for the combination of these to produce final data sets for structure solution and refinement. However, the actual data collection step is not as optimised at it could be. In particular, when the sample holder contains relatively large crystals, little or no attempt, other than cluster analysis, is made to group partial data collections from the same crystal. Conversely, in cases where there are very many crystals on the same sample holder, these may overlap producing diffraction images containing patterns from more than one crystal. This has the result either that indexing, and thus further analysis, of the diffraction images for a partial data set fails or, perhaps worse, that overlapping diffraction patterns result in erroneous estimation of diffraction spot intensities and their standard deviations. Recent developments at ESRF have attempted to address both such situations, with the program *MeshBest* allowing, during the planning of the collection of partial data sets, the recognition of areas belonging to the same crystal or the exclusion of regions containing overlapped crystals. *MeshBest* is based on the output from DOZOR (see §3.2), particularly the list of diffraction positions it produces (Figure 3.3).

In order to recognise different areas in Mesh&Collect mesh scans where diffraction images likely arise from the same crystal, diffraction spot locations on neighbouring images are compared for similarity. When they are sufficiently similar, these areas are grouped and elliptical shapes fitted to determine approximate individual crystal sizes in two dimensions (*i.e.* Figure 3.8). If desired, information on crystal size and orientation can then be used to set up individually tailored data collections for each crystal identified that optimise their diffraction potential. To recognise regions of mesh scans where diffraction images contain overlapped patterns, *MeshBest* analyses the distribution of inter-spot vectors in reciprocal space. Those images with a large number of very short vectors (shorter than the minimum expected reciprocal lattice vector) are marked as possible problematic areas and can, if desired, be excluded from subsequent data collection and/or

Table 3.3 Statistics for the GA-based merging of partial data sets collected in a Mesh&Collect experiment on small crystals of LUX. All refers to the combination of all sub data-sets collected and 01 to the combination of sub data-sets obtained using default weights in the calculation of the evaluation function. 02 and 03 summarise the combinations of sub data-sets obtained with increased weights on $\langle I/\sigma(I)\rangle$ and $CC_{1/2}$, respectively.

	All	01	02	03
No. of sub-data sets	204	45	62	77
GA Population size	—	35	35	35
GA generations	—	300	400	400
GA *R* weight	—	1	12	12
GA *I* weight	—	2.5	100	71
GA $CC_{1/2}$ weight	—		6	6
GA Completeness weight	—	2	4	4
GA groups	—	3	3	3
Resolution range (Å)				
Overall	12.52–2.80	12.51–2.80	12.52–2.80	12.52–2.80
Outer shell	2.87–2.80	2.87–2.8	2.87–2.80	2.87–2.80
Total no. of reflections				
Overall	292 976	70 083	94 175	116 756
Outer shell	16 922	4231	5574	6990
No. of unique reflections				
Overall	14 499	14 224	14 227	14 395
Outer shell	1039	928	937	994
Completeness (%)				
Inner shell	100	94.5	90.9	95.8
Outer shell	100.1	89.4	90.2	95.7
Overall	99.9	98	98	99.2
Multiplicity				
Inner shell	22.64	5.61	7.83	9.36
Outer shell	16.28	4.56	5.95	7.03
Overall	20.21	4.93	6.62	8.11
R factor (%)				
Inner shell	71	17	28.2	25.7
Outer shell	187	83.2	113.6	133.2
Overall	74.5	37	58.5	55.2
R_{meas} *(%)*				
Inner shell	72.7	18.8	30.1	27.2
Outer shell	192.6	93.5	123.4	142.2
Overall	76.1	41	62.6	58.3
$\langle I/\sigma(I)\rangle$				
Inner shell	11.78	16.26	20.10	15.44
Outer shell	3.74	5.04	5.84	4.70
Overall	7.77	9.88	11.68	9.22
$CC_{1/2}$ *(%)*				
Inner shell	93.7	95	97.4	97.3
Outer shell	68.1	59.6	67.2	74.9
Overall	94.4	94.3	91.8	95.6

processing steps. Below we present two examples of how the combination of DOZOR and *MeshBest* can be used in practice.

3.8.1 NarQ Crystals Analysed by a Mesh Scan

Several crystals of NarQ,[62] prepared in a mesophase and each of maximum dimension ~100 μm on average, were harvested on a micromesh loop which was subject to a Mesh&Collect mesh scan using a beam 30 μm in diameter at the sample position. The heat map produced by DOZOR showed several well-diffracting positions (Figure 3.14a). Subsequent analysis using *MeshBest* (Figure 3.14b) suggested several areas could be grouped together, revealing the positions of several different crystals contained on the sample support, and also revealed a number of regions containing overlapped diffraction patterns. This information was then used for optimised data collection using Mesh&Collect.

3.8.2 A 'Mishmash' of Thaumatin Crystals

A large number of thaumatin crystals ranging in size from ~10 to 40 μm in maximum dimensions were harvested on a micromesh loop. Visual inspection of the loop suggested many areas on the sample support where crystals were overlapping with each other, an impression confirmed by the appearance of the DOZOR heat map from a Mesh&Collect grid scan (Figure 3.15a). Here, analysis with *MeshBest* revealed that while the majority of the region of the sample loop contained overlapped crystals, the area on the outside of the 'crystal clump' did not. Moreover, several of the positions containing diffraction from single crystals could – as for NarQ above – be grouped as arising from a series of larger crystals (Figure 3.15b). These areas

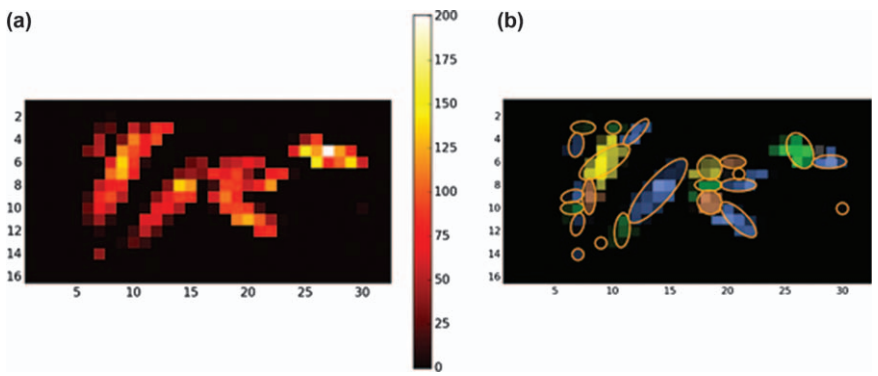


Figure 3.14 Mesh scan analysis of NarQ crystals. (a) The DOZOR heat map of the initial mesh scan. (b) the 'crystal map' calculated using MeshBest; the regions with superposed, multi-crystal diffraction patterns are depicted in grey; those regions of diffraction emanating from different single crystals are shown in different colours and fitted to elliptical 'crystal shapes' for better estimation of crystal centres and volumes.

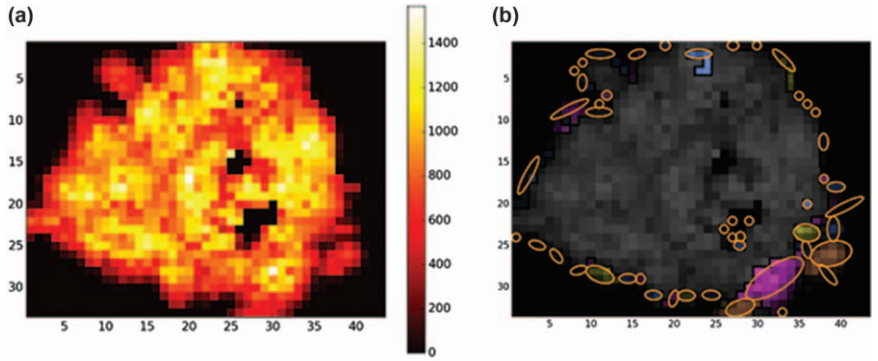


Figure 3.15 Mesh scan analysis of a clump of thaumatin crystal mounted on a solid support. (a) The DOZOR heat map of the initial mesh scan. (b) The 'crystal map' calculated using MeshBest; the regions with superposed, multi-crystal diffraction patterns are depicted in grey; those regions of diffraction emanating from different single crystals are shown in different colours and fitted to elliptical 'crystal shapes' for better estimation of crystal centres and volumes.

were then used for Mesh&Collect data collection, producing a data set of higher quality than might otherwise not have been the case.

3.9 Conclusions

Mesh&Collect provides a user-friendly pipeline for the collection and merging of partial data sets collected in multi-crystal MX experiments. Its major strengths are that crystals can be mounted on almost any type of sample support and experiments can be carried out at either cryo- or room temperature. These features make the pipeline accessible to almost any user of a synchrotron-based MX beamline. Moreover, Mesh&Collect is also widely applicable and has been used in experiments as diverse as the collection of high quality data sets from very small crystals of a membrane protein, the production of data sets for structure solution using the S-SAD, I-SAD or I-SIRAS techniques and the construction of data sets from small crystals that allow the study of ligand binding. While available on all of the ESRF's MX beamlines, Mesh&Collect is still subject to continuing evolution and improvement. Two areas (GA-based merging of partial data sets and the *MeshBest* analysis of initial grid scan) that will improve the quality of the final data sets constructed using the pipeline have been introduced here. These, when combined with other on-going developments, should result in Mesh&Collect pipelines which are even more user friendly and even more widely applicable.

Note added in proof: Since this chapter was first submitted for publication fuller details of some of the aspects discussed have been published. These include CCcluster,[63] MeshBest[64] and Mesh&Collect-based phasing using X-ray or UV RIP.[65]

References

1. H. Hope, *Acta Crystallogr., Sect. B*, 1988, **44**(Pt 1), 22–26.
2. E. F. Garman and S. Doublie, *Methods Enzymol.*, 2003, **368**, 188–216.
3. R. Henderson, *Proc. R. Soc. London, Ser. B*, 1990, **241**, 6–8.
4. G. P. Bourenkov and A. N. Popov, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2006, **62**, 58–64.
5. G. P. Bourenkov and A. N. Popov, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2010, **66**, 409–419.
6. R. L. Owen, E. Rudino-Pinera and E. F. Garman, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 4912–4917.
7. J. C. Kendrew, R. E. Dickerson, B. E. Strandberg, R. G. Hart, D. R. Davies, D. C. Phillips and V. C. Shore, *Nature*, 1960, **185**, 422–427.
8. H. N. Chapman, P. Fromme, A. Barty, T. A. White, R. A. Kirian, A. Aquila, M. S. Hunter, J. Schulz, D. P. DePonte, U. Weierstall, R. B. Doak, F. R. Maia, A. V. Martin, I. Schlichting, L. Lomb, N. Coppola, R. L. Shoeman, S. W. Epp, R. Hartmann, D. Rolles, A. Rudenko, L. Foucar, N. Kimmel, G. Weidenspointner, P. Holl, M. Liang, M. Barthelmess, C. Caleman, S. Boutet, M. J. Bogan, J. Krzywinski, C. Bostedt, S. Bajt, L. Gumprecht, B. Rudek, B. Erk, C. Schmidt, A. Homke, C. Reich, D. Pietschner, L. Struder, G. Hauser, H. Gorke, J. Ullrich, S. Herrmann, G. Schaller, F. Schopper, H. Soltau, K. U. Kuhnel, M. Messerschmidt, J. D. Bozek, S. P. Hau-Riege, M. Frank, C. Y. Hampton, R. G. Sierra, D. Starodub, G. J. Williams, J. Hajdu, N. Timneanu, M. M. Seibert, J. Andreasson, A. Rocker, O. Jonsson, M. Svenda, S. Stern, K. Nass, R. Andritschke, C. D. Schroter, F. Krasniqi, M. Bott, K. E. Schmidt, X. Wang, I. Grotjohann, J. M. Holton, T. R. Barends, R. Neutze, S. , Marchesini, R. Fromme, S. Schorb, D. Rupp, M. Adolph, T. Gorkhover, I. Andersson, H. , Hirsemann, G. Potdevin, H. Graafsma, B. Nilsson and J. C. Spence, *Nature*, 2011, **470**, 73–77.
9. F. Stellato, D. Oberthur, M. Liang, R. Bean, C. Gati, O. Yefanov, A. Barty, A. Burkhardt, P. Fischer, L. Galli, R. A. Kirian, J. Meyer, S. Panneerselvam, C. H. Yoon, F. Chervinskii, E. Speller, T. A. White, C. Betzel, A. Meents and H. N. Chapman, *IUCrJ*, 2014, **1**, 204–212.
10. P. Nogly, D. James, D. Wang, T. A. White, N. Zatsepin, A. Shilova, G. Nelson, H. Liu, L. Johansson, M. Heymann, K. Jaeger, M. Metz, C. Wickstrand, W. Wu, P. Bath, P. Berntsen, D. Oberthuer, V. Panneels, V. Cherezov, H. Chapman, G. Schertler, R. Neutze, J. Spence, I. Moraes, M. Burghammer, J. Standfuss and U. Weierstall, *IUCrJ*, 2015, **2**, 168–176.
11. S. Botha, K. Nass, T. R. M. Barends, W. Kabsch, B. Latz, F. Dworkowski, L. Foucar, E. Panepucci, M. T. Wang, R. L. Shoeman, I. Schlichting and R. B. Doak, *Acta Crystallogr., Sect. D*, 2015, **71**, 387–397.
12. C. Gati, G. Bourenkov, M. Klinge, D. Rehders, F. Stellato, D. Oberthur, O. Yefanov, B. P. Sommer, S. Mogk, M. Duszenko, C. Betzel, T. R. Schneider, H. N. Chapman and L. Redecke, *IUCrJ*, 2014, **1**, 87–94.

13. U. Zander, G. Bourenkov, A. N. Popov, D. de Sanctis, O. Svensson, A. A. McCarthy, E. Round, V. Gordeliy, C. Mueller-Dieckmann and G. A. Leonard, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2015, **71**, 2328–2343.
14. C.-Y. Huang, V. Olieric, P. Ma, E. Panepucci, K. Diederichs, M. Wang and M. Caffrey, *Acta Cryst. D*, 2015, **71**, 1399–1407.
15. A. S. Soares, J. D. Mullen, R. M. Parekh, G. S. McCarthy, C. G. Roessler, R. Jackimowicz, J. M. Skinner, A. M. Orville, M. Allaire and R. M. Sweet, *J. Synchrotron Radiat.*, 2014, **21**, 1231–1239.
16. S. Monaco, E. Gordon, M. W. Bowler, S. Delageniere, M. Guijarro, D. Spruce, O. Svensson, S. M. McSweeney, A. A. McCarthy, G. Leonard and M. H. Nanao, *J. Appl. Crystallogr.*, 2013, **46**, 804–810.
17. G. Winter, *J. Appl. Crystallogr.*, 2010, **43**, 186–190.
18. N. Coquelle, A. S. Brewster, U. Kapp, A. Shilova, B. Weinhausen, M. Burghammer and J.-P. Colletier, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2015, **71**, 1184–1196.
19. J. Foadi, P. Aller, Y. Alguel, A. Cameron, D. Axford, R. L. Owen, W. Armour, D. G. Waterman, S. Iwata and G. Evans, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2013, **69**, 1617–1632.
20. R. Giordano, R. M. Leal, G. P. Bourenkov, S. McSweeney and A. N. Popov, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2012, **68**, 649–658.
21. J. Gabadinho, A. Beteva, M. Guijarro, V. Rey-Bakaikoa, D. Spruce, M. W. Bowler, S. Brockhauser, D. Flot, E. J. Gordon, D. R. Hall, B. Lavault, A. A. McCarthy, J. McCarthy, E. Mitchell, S. Monaco, C. Mueller-Dieckmann, D. Nurizzo, R. B. Ravelli, X. Thibault, M. A. Walsh, G. A. Leonard and S. M. McSweeney, *J. Synchrotron Radiat.*, 2010, **17**, 700–707.
22. D. de Sanctis and G. Leonard, in *Notiziarao Neutroni e Luce di Sincrotrone, Consiglio Nazionale delle Ricerche*, 2014, vol. 19, pp. 24–226.
23. S. Delageniere, P. Brenchereau, L. Launer, A. W. Ashton, R. Leal, S. Veyrier, J. Gabadinho, E. J. Gordon, S. D. Jones, K. E. Levik, S. M. McSweeney, S. Monaco, M. Nanao, D. Spruce, O. Svensson, M. A. Walsh and G. A. Leonard, *Bioinformatics*, 2011, **27**, 3186–3192.
24. W. Kabsch, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2010, **66**, 125–132.
25. M. Caffrey and V. Cherezov, *Nat. Protoc.*, 2009, **4**, 706–731.
26. M. Caffrey, *Annu. Rev. Biophys.*, 2009, **38**, 29–51.
27. V. I. Gordeliy, R. Schlesinger, R. Efremov, G. Buldt and J. Heberle, *Methods Mol. Biol.*, 2003, **228**, 305–316.
28. A. Vagin and A. Teplyakov, *J. Appl. Crystallogr.*, 1997, **30**, 1022–1025.
29. V. I. Borshchevskiy, E. S. Round, A. N. Popov, G. Buldt and V. I. Gordeliy, *J. Mol. Biol.*, 2011, **409**, 813–825.
30. G. N. Murshudov, P. Skubak, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long and A. A. Vagin, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2011, **67**, 355–367.
31. P. Emsley, B. Lohkamp, W. G. Scott and K. Cowtan, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2010, **66**, 486–501.

32. W. Brehm and K. Diederichs, *Acta Crystallogr., Sect. D*, 2014, **70**, 101–109.
33. T. O. Yeates, *Method Enzymol.*, 1997, **276**, 344–358.
34. C. W. Murray and T. L. Blundell, *Curr. Opin. Struct. Biol.*, 2010, **20**, 497–507.
35. C. W. Murray, M. G. Carr, O. Callaghan, G. Chessari, M. Congreve, S. Cowan, J. E. Coyle, R. Downham, E. Figueroa, M. Frederickson, B. Graham, R. McMenamin, M. A. O'Brien, S. Patel, T. R. Phillips, G. Williams, A. J. Woodhead and A. J. Woolford, *J. Med. Chem.*, 2010, **53**, 5942–5955.
36. L. Oster, S. Tapani, Y. Xue and H. Kack, *Drug Discovery Today*, 2015, **20**, 1104–1111.
37. K. Cole, C. G. Roessler, E. A. Mule, E. J. Benson-Xu, J. D. Mullen, B. A. Le, A. M. Tieman, C. Birone, M. Brown, J. Hernandez, S. Neff, D. Williams, M. Allaire, A. M. Orville, R. M. Sweet and A. S. Soares, *PLoS One*, 2014, **9**, e101036.
38. D. de Sanctis, A. Beteva, H. Caserotto, F. Dobias, J. Gabadinho, T. Giraud, A. Gobbo, M. Guijarro, M. Lentini, B. Lavault, T. Mairs, S. McSweeney, S. Petitdemange, V. Rey-Bakaikoa, J. Surr, P. Theveneau, G. A. Leonard and C. Mueller-Dieckmann, *J. Synchrotron Radiat.*, 2012, **19**, 455–461.
39. F. Cipriani, M. Rower, C. Landret, U. Zander, F. Felisaz and J. A. Marquez, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2012, **68**, 1393–1399.
40. Z. Dauter, M. Dauter, E. de La Fortelle, G. Bricogne and G. M. Sheldrick, *J. Mol. Biol.*, 1999, **289**, 83–92.
41. E. Micossi, W. N. Hunter and G. A. Leonard, *Acta Crystallogr., Sect. D*, 2002, **58**, 21–28.
42. J. Abendroth, A. S. Gardberg, J. I. Robinson, J. S. Christensen, B. L. Staker, P. J. Myler, L. J. Stewart and T. E. Edwards, *J. Struct. Funct. Genomics*, 2011, **12**, 83–95.
43. T. Nakane, S. Hanashima, M. Suzuki, H. Saiki, T. Hayashi, K. Kakinouchi, S. Sugiyama, S. Kawatake, S. Matsuoka, N. Matsumori, E. Nango, J. Kobayashi, T. Shimamura, K. Kimura, C. Mori, N. Kunishima, M. Sugahara, Y. Takakyu, S. Inoue, T. Masuda, T. Hosaka, K. Tono, Y. Joti, T. Kameshima, T. Hatsui, M. Yabashi, T. Inoue, O. Nureki, S. Iwata, M. Murata and E. Mizohata, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, 13039–13044.
44. J. P. Rose, B. C. Wang and M. S. Weiss, *IUCrJ*, 2015, **2**, 431–440.
45. T. Weinert, V. Olieric, S. Waltersperger, E. Panepucci, L. Chen, H. Zhang, D. Zhou, J. Rose, A. Ebihara, S. Kuramitsu, D. Li, N. Howe, G. Schnapp, A. Pautsch, K. Bargsten, A. E. Prota, P. Surana, J. Kottur, D. T. Nair, F. Basilico, V. Cecatiello, S. Pasqualato, A. Boland, O. Weichenrieder, B. C. Wang, M. O. Steinmetz, M. Caffrey and M. Wang, *Nat. Methods*, 2015, **12**, 131–133.
46. A. Gonzalez, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2003, **59**, 1935–1942.

47. D. de Sanctis, M. Oscarsson, A. Popov, O. Svensson and G. Leonard, *Acta Crystallogr., Sect. D: Struct. Biol.*, 2016, **72**, 413–420.
48. P. Skubak and N. S. Pannu, *Nat. Commun.*, 2013, **4**, 2777.
49. I. Melnikov, V. Polovinkin, K. Kovalev, V. Shevchenko, M. Shevtsov, V. Borshchevskiy, V. Cherezov, G. Leonard, V. Gordeliy and A. Popov, *Acta Crystallogr., Sect. A*, 2016, **72**, s199.
50. I. Gushchin, V. Shevchenko, V. Polovinkin, K. Kovalev, A. Alekseev, E. Round, V. Borshchevskiy, T. Balandin, A. Popov, T. Gensch, C. Fahlke, C. Bamann, D. Willbold, G. Buldt, E. Bamberg and V. Gordeliy, *Nat. Struct. Mol. Biol.*, 2015, **22**, 390–395.
51. I. Gushchin, V. Shevchenko, V. Polovinkin, V. Borshchevskiy, P. Buslaev, E. Bamberg and V. Gordeliy, *FEBS J.*, 2016, **283**, 1232–1238.
52. D. Nurizzo, T. Mairs, M. Guijarro, V. Rey, J. Meyer, P. Fajardo, J. Chavanne, J. C. Biasci, S. McSweeney and E. Mitchell, *J. Synchrotron Radiat.*, 2006, **13**, 227–238.
53. G. M. Sheldrick, *Acta Crystallogr., Sect. A*, 2008, **64**, 112–122.
54. C. Mueller-Dieckmann, M. W. Bowler, P. Carpentier, D. Flot, A. A. McCarthy, M. H. Nanao, D. Nurizzo, P. Pernot, A. Popov, A. Round, A. Royant, D. de Sanctis, D. von Stetten and G. A. Leonard, *Eur. Phys. J. Plus*, 2015, **130**, 1–11.
55. R. W. Grosse-Kunstleve, N. K. Sauter, N. W. Moriarty and P. D. Adams, *J. Appl. Crystallogr.*, 2002, **35**, 126–136.
56. U. Zander, M. Cianci, N. Foos, C. S. Silva, L. Mazzei, C. Zubieta, A. de Maria and M. H. Nanao, *Acta Crystallogr., Sect. D: Struct. Biol.*, 2016, **72**, 1026–1035.
57. R. B. Ravelli, H. K. Leiros, B. Pan, M. Caffrey and S. McSweeney, *Structure*, 2003, **11**, 217–224.
58. D. de Sanctis and M. H. Nanao, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2012, **68**, 1152–1162.
59. P. A. Karplus and K. Diederichs, *Science*, 2012, **336**, 1030–1033.
60. K. Diederichs and P. A. Karplus, *Nat. Struct. Biol.*, 1997, **4**, 269–275.
61. M. S. Weiss, *J. Appl. Crystallogr.*, 2001, **34**, 130–135.
62. V. Stewart, *Biochem. Soc. Trans.*, 2003, **31**, 1–10.
63. G. Santoni, U. Zander, C. Mueller-Dieckmann, G. Leonard and A. Popov, *J. Appl. Crystallogr.*, 2017, **50**, 1844–1851.
64. I. Melnikov, O. Svensson, G. Bourenkov, G. Leonard and A. Popov, *Acta Crystallogr., Sect. D*, 2018, **74**, DOI: 10.1107/S2059798318002735.
65. N. Foos, C. Seuring, R. Schubert, A. Burkhardt, O. Svensson, A. Meents, H. Chapman and M. H. Nanao, *Acta Crystallogr., Sect. D*, 2018, **74**, DOI: 10.1107/S2059798318001535.

CHAPTER 4

Radiation Damage in Macromolecular Crystallography

ELSPETH F. GARMAN*[a] AND MARTIN WEIK*[b]

[a] Laboratory of Molecular Biophysics, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK; [b] Institut de Biologie Structurale, Univ. Grenoble Alpes, CEA, CNRS, F-38044 Grenoble, France
*Email: elspeth.garman@bioch.ox.ac.uk; weik@ibs.fr

4.1 Introduction

The success of macromolecular structure solution by X-ray crystallographic methods relies firstly on the accurate measurement of the Bragg reflection intensities and secondly on the determination of the related phases (the so called 'phase problem'). The former requirement is challenged by the radiation damage induced by X-rays absorbed in the crystal during the diffraction experiment, an unavoidable by-product of the irradiation of matter. As noted by Blake and Philips as early as 1962 when studying the progression of damage in room temperature (RT) crystals of myoglobin, reflection intensities changed during the course of the experiment, and while the overall diffracting power of the crystal always decreases as a function of irradiation time, individual reflection intensities can increase as well as decrease. They also noted that the damage might be structurally specific, *i.e.*, particular amino acids were damaged before others, but since at that time the primary sequence of myoglobin was unknown, they were not able to identify which

Chemical Biology No. 8
Protein Crystallography: Challenges and Practical Solutions
Edited by Konstantinos Beis and Gwyndaf Evans
© The Royal Society of Chemistry 2018
Published by the Royal Society of Chemistry, www.rsc.org

residues were most affected. To monitor the health of their samples, pioneering protein crystallographers working with RT samples and using single reflection detectors measured the intensity of a reference reflection by returning to it regularly. When its original intensity (I_1) had reduced to $0.85I_1$, they would discard that crystal and mount a new sample, although if there was a shortage of crystals they would relax this criterion to $0.7I_1$.[1]

In the late 1980s, small molecule cryo-techniques were imported into Macromolecular Crystallography (MX)[2] in efforts to reduce radiation damage rates to ribosome crystals, and then these methods developed apace for MX when Teng[3] suggested the use of a wire loop (soon replaced, by a number of other research groups, by various types of fibre loop having very low diffracting power). This loop held the crystal in the surface tension of a cryoprotecting buffer (used to avoid the formation of ordered ice) before it was flash-cooled to 77 K/~100 K in liquid/gaseous nitrogen respectively, and data were then collected at around 100 K. Systematic cryo-crystallographic techniques for MX were subsequently developed and widely disseminated.[4–7] The concomitant introduction of an easy to use open-flow cryostat, with which crystals could reliably be kept at 100 K during the diffraction experiments,[8] resulted in the widespread adoption of loop-mounted cryo-methods by the MX community. Currently, of the MX structures deposited in the Protein Data Bank (PDB) with specified temperature for data collection, 86% are reported to have had the crystal held below 105 K (91% held below 160 K) for data collection.

Although cryo-cooling reduces the rate of radiation damage by around a factor of 70 compared with that seen at RT,[9] as early as 1994 damage with the crystal held at 100 K was noted when using a white beam at a 2nd generation synchrotron source.[10] With the advent of the much higher flux density (photons $mm^{-2}\,s^{-1}$) X-ray beams provided by 3rd generation facilities, firstly at the European Synchrotron Radiation Facility (ESRF) in Grenoble, observations of radiation damage effects at 100 K when using monochromatic beams became more frequent. The first three studies focusing on the associated effects were published in 2000.[11–13] Over the last 17 years, a wide variety of studies have been conducted to investigate the physical and chemical factors affecting the rates of radiation damage during MX, and there is now a significant body of literature on the subject. Over 70 papers have been published in eight special issues of the *Journal of Synchrotron Radiation* (Nov. 2002, May 2005, Jan. 2007, March 2009, May 2011, Jan. 2013, March 2015, Jan. 2017) resulting from a series of International Workshops on radiation damage to crystalline biological samples held over the last 18 years, and the interested reader can find in them a source of much more detail and many further references. Below we summarise the main results from these and other investigations.

It is worth noting here that radiation damage has become a major issue in MX, and will be exacerbated by the even greater flux densities becoming available at the 4th generation synchrotrons now coming on line. At these, the diffraction lifetime of a protein crystal held at 100 K will be merely

milliseconds, and datasets will routinely be assembled from images collected from a large number of crystals (serial or multi-crystal data collection schemes). Serial crystallography data collection schemes were initially developed for experiments using the femtosecond (10 to 80 fs) pulses produced by X-ray Free Electron Laser (XFEL) facilities, in which damage-free 'diffraction before destruction'[14] RT structures of proteins can be obtained from one-shot-per-crystal experiments (so-called serial femtosecond crystallography, SFX).[15,16]

4.2 How Do X-ray Photons Interact with Matter?

At the incident beam energies commonly used for MX, X-ray photons interact in three main physical ways with the matter in the sample, and these are summarized in Figure 4.1, as well as in detail in Nave 1995.[17]

Firstly, the X-ray can elastically scatter from an atom in the sample (Figure 4.1(a)) and this is the desired interaction for diffraction as, if the Bragg condition is met, it can then constructively interfere with other elastically scattered X-rays to contribute to a reflection. Secondly, the X-ray can undergo inelastic scattering (Compton effect, Figure 4.1(b)), leaving some of its energy in the material. This interaction results in a lower energy outgoing photon and the excitation or ionization of the atom, the latter giving a low energy electron carrying the residual kinetic energy. The third process is the photoelectric effect (Figure 4.1(c)) in which the incident X-ray photon is completely absorbed by the atom with the concomitant ejection of a high energy 'photoelectron' which carries the same energy as the incident photon minus the atomic binding energy of that electron. The atom is left in an excited state, and can de-excite by one of two processes: by ejecting an Auger electron with the accompanying change in the energy of another electron as it drops into the initially vacated hole, or by emitting a fluorescent X-ray. The latter process becomes more probable with increasing atomic number; it is negligible for all ''organic'' elements but rises to 30% for iron.

At 100 K, the so-called 'primary' photoelectron can penetrate the sample by up to several microns (deduced experimentally for 18.7 keV incident X-rays to have a range of 3–4 μm[18]). During their passage through the sample the liberated photoelectrons can excite many other atoms which can, in turn, release secondary electrons; a photoelectron ejected by a 12.4 keV (1 Å) incident beam can cause up to ~500 further ionisations, assuming 25 eV is required for each of these subsequent events.[19] For a 100 μm thick metal-free crystal, only 2% of an $E_{inc} = 12.4$ keV beam will interact at all, and as shown in Figure 4.1, the three processes mentioned above (elastic scattering, Compton and photoelectric effects) account for 8%, 8% and 84% of the interaction, respectively.[20] Therefore, the diffracted X-rays only constitute 0.16% of the incident beam, and the remaining 1.84% of the incident beam that interacts is responsible for the radiation damage effects, giving a ratio of over 10 potentially damaging interactions for every diffracted photon.

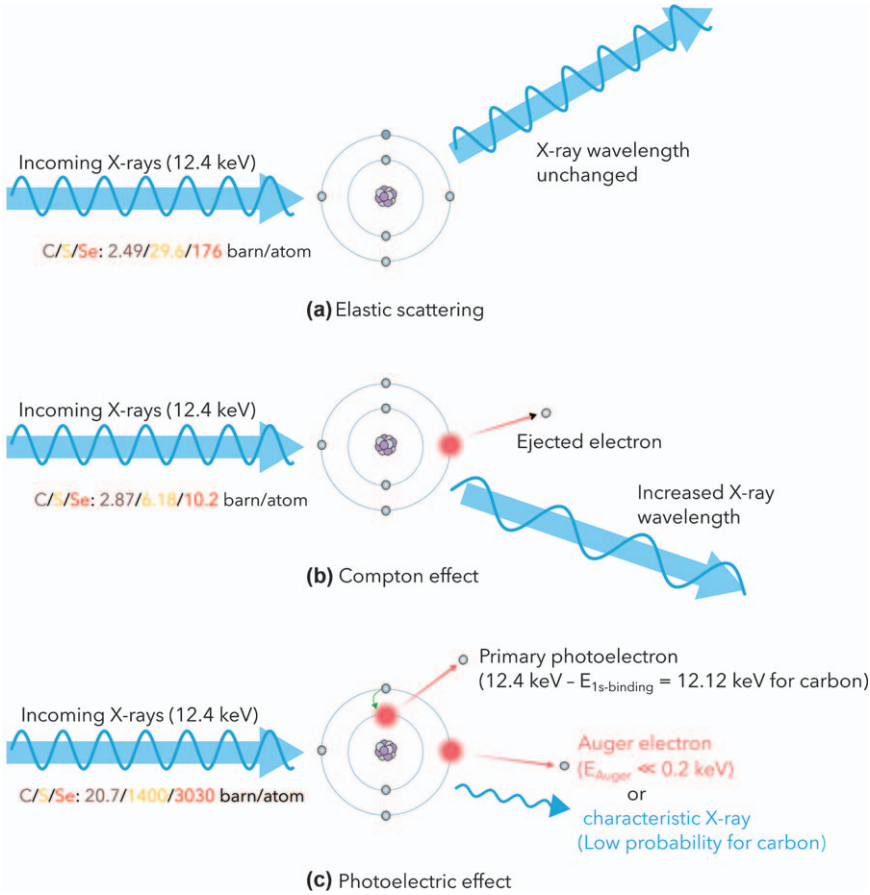


Figure 4.1 The three possible interaction processes of an incident X-ray photon with atoms in a crystal. (a) Elastic (Thomson) scattering (gives the diffraction pattern), (b) Compton effect and (c) the photoelectric effect. The photon cross-sections for each interaction (in units of barn/atom: 1 barn = 10^{-28} m^2) are given for C, S and Se atoms, with an incident 12.4 keV (1 Å) X-ray beam, as specified in the XCOM Photon Cross Section Database (http://physics.nist.gov/PhysRefData/Xcom/html/xcom1.html).

The success of cryo-temperature data collection[7] in reducing the rate of radiation damage is due to the fact that at 100 K, the diffusion rates of almost all secondary radicals produced by the subsequent excitation events, are significantly lower than at RT. For example, below 110 K, one of the two predominant radicals produced by the radiolysis of the omnipresent water, the hydroxyl radical, $OH^{\bullet}$, is thought to be immobile.[21] However, the other main radical formed by water radiolysis is the electron (which will rapidly convert to a hydrogen atom under very acidic conditions). At all temperatures these electrons, as well as hydrogen atoms and holes (electron-loss centres), are able to migrate by quantum mechanical tunneling,[22] and can

move around a protein structure even when the crystal is held at 100 K. Some of the species created locally by absorption events will, of course, recombine, and this process competes with charge migration.

The absorption events described above can be divided into 'direct' (occurring in the protein) and 'indirect' (occurring in the solvent including the hydration shell around the protein molecule). They are also categorized as 'primary' (usually the event which liberates the initial photoelectron) and 'secondary' (usually all the subsequent species induced by the energy lost in the sample by 'primary' photoelectron), though this terminology has not been uniformly applied in the literature.

4.3 Global and Specific Radiation Damage Effects at 100 K and Below

There are two predominant classes of radiation damage effects detected in crystals held at 100 K for data collection, one evidenced in reciprocal space and the other in real space.

The former is usually designated 'global damage' and is visible on the diffraction pattern as a gradual fading of the reflection intensity, with the highest-resolution reflections being the first to disappear.[10] This effect often limits the resolution to which data can be integrated and consequently compromises the atomic detail of the resulting structure. A more subtle effect is that unit cell volume increases with exposure,[12,13,23] which means that as the experiment proceeds, the crystal becomes non-isomorphous with itself. This unit cell volume expansion was initially suggested as a possible metric for monitoring radiation damage progression in reciprocal space,[12] but systematic experimental studies on crystals of a number of different proteins showed it to be too unreproducible even using parts broken from the same large crystal.[24,25] The effect is now thought to be due to the buildup of hydrogen gas which has become trapped at domain boundaries.[26] Its source is the radiolytic reactions proceeding in the crystal from the processes described above.

Further effects observed in reciprocal space as the exposure time lengthens are that the Wilson *B*-factors increase, the merging *R*-factors increase, the $I/\sigma(I)$ values decrease since the overall diffraction intensity decreases and the noise ($\sigma(I)$) concomitantly increases, and finally, crystal mosaicity usually increases.

Some of these changing parameters can be used to track the damage rates after (and sometimes during) the experiment, and there are two metrics that are commonly used for this purpose: firstly the summed intensity of a dataset or data wedge (I_n) divided by the initial summed intensity of that dataset or wedge (I_1), I_n/I_1[27] as shown in Figure 4.2(a), and secondly the relative scaling *B*-factor, B_{rel} (Figure 4.2(b)). This is the difference in Wilson *B*-factor between the *n*th dataset or wedge and that of the first dataset or wedge.[28] These parameters are best plotted against the dose, *D*, which is

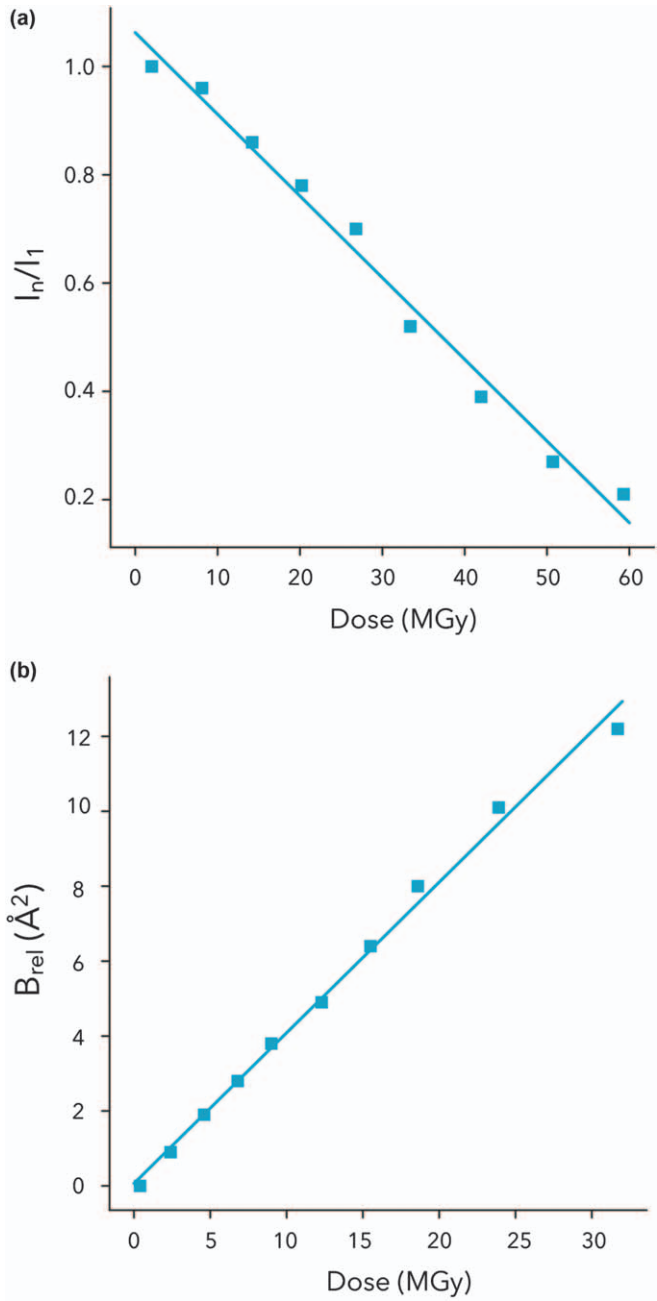


Figure 4.2 (a) Normalised summed intensity decay of complete consecutive datasets for a holoferritin crystal irradiated at 100 K, I_n/I_1, against dose. (b) Relative *B*-factor, $B_{rel} = B_n - B_1$ for the same holoferritin crystal. Plots show a linear decay and increase, respectively, with increasing dose over the reported dose range (fitted lines shown). Data collected as reported in ref. 27.

defined as the energy absorbed per unit of mass of sample ($J\,kg^{-1} = Gy$, gray). The estimation of dose is discussed detail in 4.4 below. A plot of B_{rel} against dose usually shows linear behaviour, and the gradient for a particular crystal species provides a measure of its radiation sensitivity. A coefficient of sensitivity can be defined as $s_{AD} = \Delta B_{rel}/8\pi^2\Delta D$ and this is $\sim$0.012 $Å^2\,MGy^{-1}$ at 100 K for chicken egg-white lysozyme (CEWL) crystals. The s_{AD} value has been found to be relatively constant for CEWL, thaumatin, catalase and apoferritin crystals.[28]

A third occasionally-used metric is the pairwise *R*-factor, R_d, between identical and symmetry related reflections occurring on different diffraction images, and for this parameter, the *x*-axis is the difference in dose ΔD between those images[29] rather than just the dose value at which they were detected.

Conclusions drawn from plots of I_n/I_1, B_{rel}, and R_d extracted from the same diffraction data do not usually concur. An example of this disagreement can be found in attempts to determine the efficacy at 100 K of $NaNO_3$, an electron scavenger, on the dose lifetime of CEWL crystals when they were soaked for various times (see Figures 3 and 4 of De la Mora *et al.*,[30]). The three metrics give different results for which is the most effective scavenger soak time. This is somewhat unsatisfactory and a consensus is yet to emerge as to which is the most reliable one for routine use. The underlying cause for the disagreement in results given by these metrics is not understood.

The second major class of radiation damage observed in MX manifests itself in real space electron density maps as X-ray induced changes occurring preferentially to particular amino acids, so-called 'specific structural damage', in a well-established and reproducible order. As mentioned above, this was first noted as likely at RT by Blake and Phillips in 1962, although they thought that it was leucines that were most affected as they had no sequence information (or structure). However, the high susceptibility of disulfide bonds to damage at RT was reported in 1988,[31] and in the year 2000 broken disulfide bonds (Figure 4.3(a)) were also observed in three different systematic experiments at 100 K, as was the clear decarboxylation of aspartic and glutamic acid residues[11–13] (Figure 4.3(b)). The radiation chemistry mechanisms involve reduction of the disulfide bonds, whereby the sulfurs capture an electron, the bonds elongate and then break following a complex multi-track model.[32] They form a disulfide radical anion,[33] whose signature absorption peak at 400 nm has been detected using both off-line[24] and on-line microspectrophotometers.[34] Conversely, the decarboxylation involves an oxidation reaction.[35] Damage has also been seen on methionine residues with difference electron density appearing around the C–S bond (Figure 4.3(c)).

If specific damage was solely due to primary interaction events, it would be expected that it would occur first to the residues containing atoms with the greatest X-ray cross-sections (the sulfurs in cysteines and methionines). Although disulfides are very susceptible, cysteines are not, and glutamates and aspartates are more sensitive than methionines. From this it can be

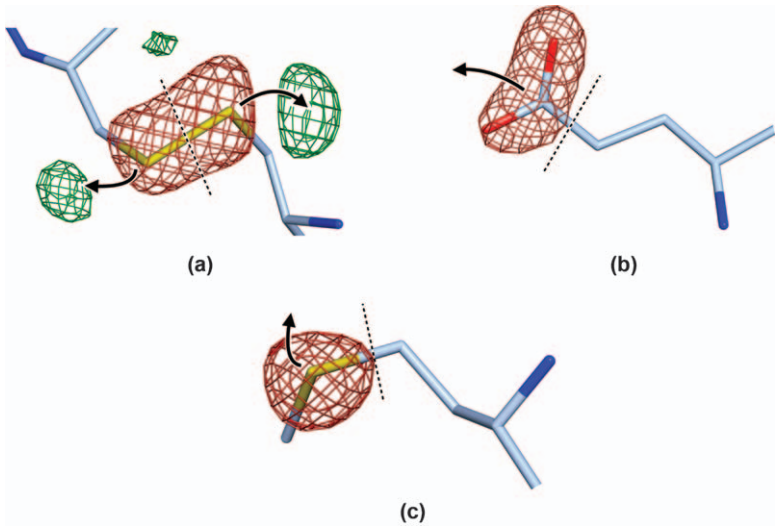


Figure 4.3 Specific structural damage showing (a) a disulfide bond with the new sulfur positions after cleavage (green density), (b) a glutamate decarboxylation, and (c) a methionine sulfur disordering in myrosinase, PDB: 1DWA.[13] Fourier difference maps between datasets 1 and 4 (dose ~14 MGy) collected from the same crystal, $F_{obs}(4) - F_{obs}(1)$, calculated using the phases of the structure refined from dataset 1. Red difference density indicates disordering of the atomic positions.

concluded that specific damage is caused by the secondary radicals that are produced by the primary absorption event[23,24] and are able to tunnel through the molecule. Note that all amino acids of the same sort in a protein structure are not damaged at the same rate, which is further evidence for specific damage being a secondary rather than a primary event. The reasons for the observed differential rates are not yet understood, although several studies have investigated various physicochemical parameters to find correlations.[36,37]

In the MX radiation damage literature, cleavage of the OH group from tyrosine residues is often mentioned as the third most likely specific damage event to occur, due to an early report that tentatively suggested this was the possible cause of difference density observed round the oxygen of a tyrosine residue.[13] However, a very recent reanalysis of these original data, along with a survey of all the damage series deposited in the PDB, has concluded that OH cleavage does not occur.[38] This conclusion is in accord with a wide body of aqueous solution radiation chemistry literature which has not reported the observation of this energetically unfavourable event.

An additional, and very important, comment regarding specific damage is that protein active site residues seem to be the most susceptible part of a protein, possibly due to the geometric and/or chemical strain they may be under in order to perform their functions.[12,39] For instance, observations of active site damage have been reported in acetylcholinesterase,[11]

bacteriorhodopsin,[40] photoactive yellow protein,[41] DNA photolyase,[42] malate dehydrogenases,[36] carbonic anhydrase[43] and fluorescent proteins.[44]

Specific structural damage instances must be clearly distinguished from changes related to enzyme mechanism, since they could cause misleading biological conclusions to be extracted from a protein structure. For instance there has been focused efforts to separate the intermediate states of the bacteriorhodopsin photocycle from radiation-induced structural and chemical modifications,[45] and robust mechanistic conclusions have now been drawn from a careful clustering analysis of the available PDB structures.[46]

The specific damage suffered by proteins is now well characterized but far less was known concerning radiation-induced effects on protein-nucleotide complexes. Two recent investigations have shed light on the relative damage rates to DNA and RNA as compared with protein through systematic studies of a protein–DNA complex[47] and a 91 kDa RNA–protein complex.[48] These works showed conclusively that for the complexes studied, both DNA and RNA suffer significantly slower rates of site specific damage than does protein.

In a standard cryo-crystallographic experiment, the sample is held at 100 K. Neither global nor specific damage is significantly reduced when the sample temperature is kept below 100 K,[26,49,50] with metal reduction being an important exception (see below). Consequently, helium cooling has not replaced nitrogen cooling as the routinely used cryogen.

4.4 Estimating the Absorbed Dose and Dose Limits

As mentioned above, the reciprocal space variables I_n/I_1, B_{rel}, and R_{d} can be plotted for the diffraction data in order to monitor radiation damage progression. The most useful generally applicable x-axis metric against which to plot these quantities is the absorbed dose, the energy absorbed per unit mass of sample irradiated during the measurements. This is much more fruitful than using image number, time, or incident cumulative flux for the x-axis, since dose is a quantity that can be estimated from the experimental conditions, allowing comparison of studies from different sites and researchers.

However, the calculation of the absorbed dose, which is not a measurable variable, is non-trivial and requires knowledge of a number of parameters pertaining to the particular beam used and the crystal being irradiated. The incident beam energy, flux, size (if collimated by slits or an aperture) and profile in the horizontal and vertical planes is needed, as well as the exposure time. The size and atomic composition of the crystal (number of amino acids, DNA or RNA bases, endogenous heavy atoms and those in the solvent) must also be estimated in order for the X-ray absorption coefficient, μ_{abs}, to be calculated, as this governs the proportion of the incident beam that will be absorbed. Since they have much larger X-ray cross-sections than the organic elements, the presence of a few heavy atoms either in the protein

or in the solvent can have a significant impact on the absorption coefficient, so they must be included in the computation.

Access to the beam profile and flux parameters requires synchrotron MX beamlines to be regularly calibrated, and protocols for this have been established[51] and are carried out routinely at some sites. If an experimenter wants to estimate the dose, the beamline scientist should be informed of this beforehand, so that, if necessary, the measurements can be made at setup time.

As proposed by Henderson,[52] through analogy with his observations of the lifetime of 2-D crystals held at 77 K in electron diffraction experiments, there is believed to be a maximum dose that can be tolerated by a biological crystal before its diffraction intensity is halved, called the 'half-dose', $D_{1/2}$. Making a number of assumptions, particularly about electron penetration depths, he estimated that for MX, $D_{1/2}$ could be expected to be ~20 MGy at 100 K. This 'Henderson limit' has now been experimentally measured for MX at 100 K and found to be $D_{1/2} = 43$ (± 3) MGy.[27] Data to a resolution of 2.2 Å were collected from four holo-ferritin (contains one iron atom for every two amino acids) and three apo-ferritin (no iron at all) crystals and the intensity decay was monitored as a function of dose. Since the absorption coefficients of these two crystal types differ by more than a factor of 2 (μ_{abs} at 13.2 keV = 1.43 and 0.67 mm^{-1}, respectively), this was a test of the pivotal basic assumption, proposed by Blake and Phillips for RT samples, that the rate of radiation damage is proportional to dose. To ensure that the crystal heavy atom content was well determined for the dose calculations, the elemental composition of the holo- and apo-ferritin crystals was measured using microbeam particle induced X-ray emission analysis.[53] The two different crystal types showed remarkably consistent rates of diffraction intensity decay with dose, but analysis of the electron density maps derived from successive datasets showed that by the time the crystal had reached half its initial diffracting power, the protein structures had suffered significant specific structural damage. Therefore, a more conservative limit for the dose that could be absorbed was proposed, $0.7I_1$ as opposed to $0.5I_1$, which corresponded to an absorbed dose of 30 MGy.[27] Intriguingly, this $0.7I_1$ limit for the intensity lifetime of a protein crystal at 100 K was determined independently from, but concurs with, the RT limit of $0.7I_1$ mentioned above as having been recommended by Blundell and Johnson.[1]

Other proposals have been made regarding a dose limit for cryo-MX, among them a resolution-dependent limit of 10 MGy $Å^{-1}$;[54] for example a 2.5 Å diffraction pattern would be expected to fade to 3.5 Å after the sample had absorbed 10 MGy. An equivalent RT value has yet to be established since a large range of doses to reach $0.7I_1$ have been documented (from a few kGy to ~1.5 MGy), and a value of $D_{1/2} = 150$ kGy at RT has recently been published.[55]

It must be stressed that these are *limits* and are not guaranteed doses that can safely be absorbed by crystals in MX experiments. Therefore, no more than half of the crystal diffraction intensity is expected to survive beyond the

limit, but many crystals at 100 K have a lower $D_{1/2}$ than 43 MGy (*e.g.* CEWL $D_{1/2}$ ~10 MGy[23,30]). This is mainly because the estimation of absorbed dose accounts for the physics of the experiment, but not for the radiation chemistry effects. For instance, the crystal might have a susceptible residue such as a glutamic or aspartic acid at a crystal contact, which becomes specifically damaged and consequently causes lattice disordering, a situation which would result in faster rates of global damage.[56]

In a standard MX experiment, currently a typical dose rate for a 100 μm cubed metal-free crystal is around 1 MGy s^{-1} in a 100 μm×100 μm and 10^{13} photons s^{-1} beam of 12.4 keV (1 Å) X-rays. Since each absorbed X-ray can induce the release of up to 500 electrons (estimated by taking the average energy loss to produce an electron from the valence shell of a constituent atom to be ~25 eV), it can be computed that a dose of 1 MGy will induce approximately one ionisation per 20 amino acid residues in the crystal.[19]

Since estimation of the absorbed dose in an MX experiment is fairly involved, a convenient computer code for its calculation called RADDOSE-3D[57] has been developed over the last few years, and is available to the community on a Web server at www.raddo.se and for download at https://github.com/GarmanGroup/RADDOSE-3D. It has replaced its predecessor, RADDOSE (Murray *et al.*, 2004), which went through three versions of increasing sophistication. Using the estimated absorption coefficients and tabulated X-ray cross-sections, RADDOSE computed the attenuation of the beam as it traversed the crystal to arrive at the amount transmitted, which was then subtracted from the incident flux to arrive at the number of photons remaining in the crystal, hence giving the total energy absorbed. For the convenience of the user, buffer concentrations could be entered as mM to account for their heavy atom content. RADDOSE version 2 additionally took into account the possible escape of fluorescent photons originating from the de-excitation of heavier atoms following photoelectron ejection[58] (see Figure 4.3(c)), and version 3 included the energy lost by low energy electrons resulting from the Compton effect (see Figure 4.3(b)), neglected in versions 1 and 2. The difference in dose from this effect is negligible below incident X-ray energies of 20 keV and only becomes significant above about 40 keV when the Compton cross-sections increase.[59] The dose returned by all three versions was the maximum absorbed anywhere in the crystal.

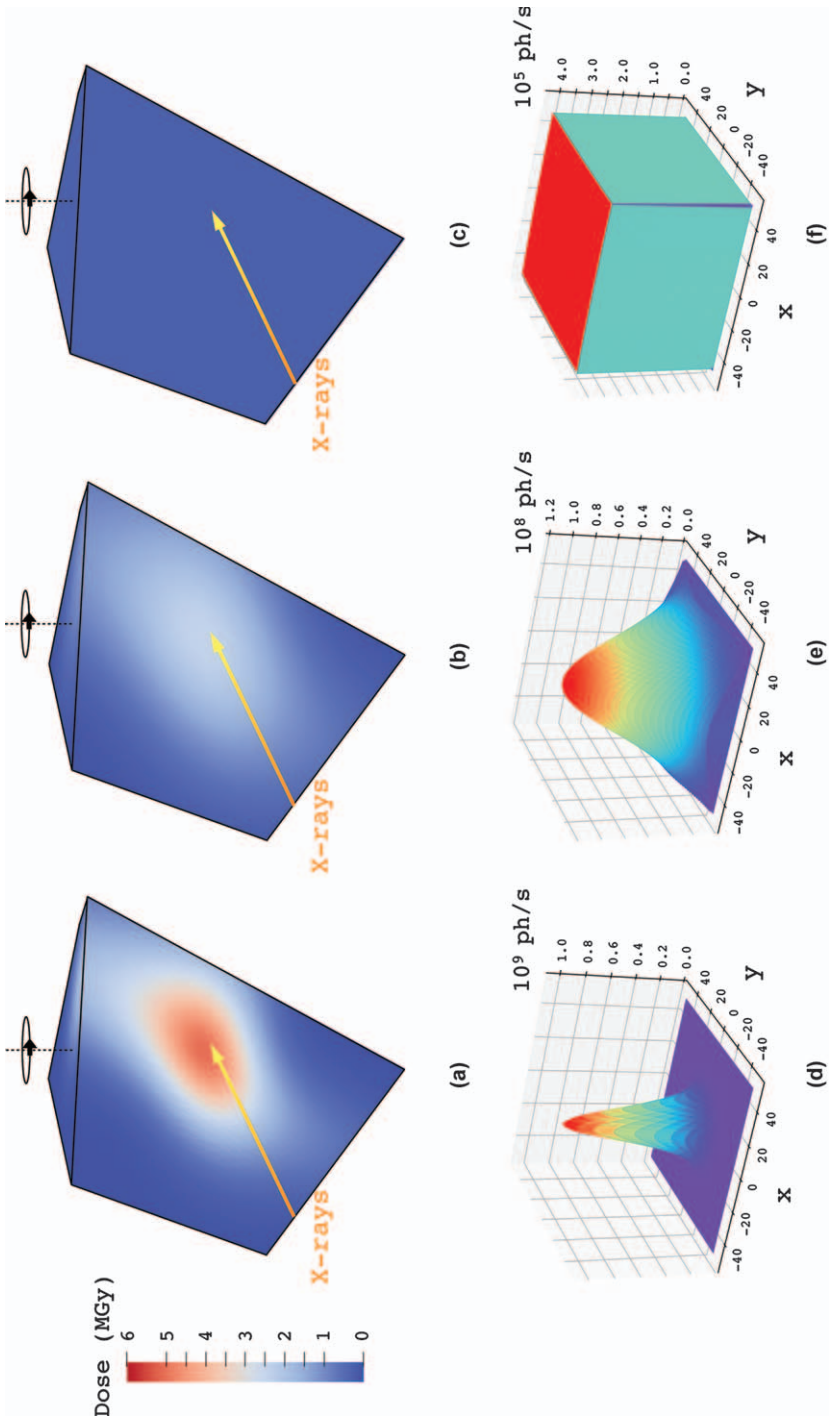
RADDOSE was developed when, in general, X-ray beams for MX were bigger than the crystals being irradiated, so that crystal rotation did not have to be considered; it was enough to know the exposure time, as the crystal was bathed completely in the beam. However, with the advent of much smaller X-ray beams (*e.g.* beamline I24 at Diamond Light Source which currently delivers a 6 μm (horizontal)×9 μm (vertical) Full Width Half Maximum (FWHM) Gaussian shaped beam) in which, for any crystal larger than the beam, fresh un-irradiated volume is being continuously rotated into the beam, a new approach to dose estimation was required. Hence RADDOSE-3D[57] was developed to provide spatial and temporal dose field modelling for a wide range of complex MX data collection scenarios,[60] such

as for helical scans[61] designed to spread the dose over the available crystal volume, and for multi-wavelength anomalous dispersion (MAD) phasing experiments.

Given the RADDOSE-3D modelling program, it is possible to change various input parameters to probe their effect on the dose distributions. Examples of the results of this capability are shown in Figure 4.4 for a $(100\ \mu m)^3$ cuboid crystal irradiated with two different FWHM Gaussian profile beams ($20\times20\ \mu m^2$ and $60\times60\ \mu m^2$) and also for a uniform top-hat profile beam. It can be seen that the dose distributions, and in particular the maximum dose, vary from 6 MGy for the small beam to less than 1 MGy for the top-hat beam for the same total incident flux. Dose distributions having a large range of values can have an impact on the resulting data quality[62] and electron density maps, since both the unit cell volume and the degree of specific damage are a function of dose. For Gaussian profile beams, molecules at the centre of the crystal remain in the hottest part of the beam throughout the rotation and will be damaged much more than those irradiated by the tails. Thus, a range of structural states are being probed at any instant, whereas for the top-hat profile beam, the crystal is homogeneously irradiated and all the protein molecules should be in a similar state at any one time.

A systematic study using 43 datasets collected from 34 different crystals revealed that the differences in damage rates could only be explained in terms of correlation with the beam profiles used for the particular data collections.[63] It is also worth noting that the cause of the disappointing success rates achieved in early MAD experiments at 100 K carried out at the first tunable MX beamlines was eventually traced to inhomogeneous beam profiles. A recent analysis of the effect of increasing multiplicity and radiation damage on the quality of the sulfur anomalous signal, from five different thaumatin crystals irradiated in a homogeneous top-hat beam, has been carried out. The results led the researchers to propose an empirical rule which uses a statistic derived from the width of the normalized distribution of a set of anomalous structure factor differences to identify the dose (*point of diminishing returns*) at which the gains in data quality due to increased multiplicity are outbalanced by the losses due to decreases in signal to noise caused by radiation damage. This rule allows data collection parameters for single wavelength anomalous dispersion (SAD) experiments to be optimised and the best data for the structure determination to be selected. The *point of diminishing returns* for data collection on thaumatin crystals at 100 K is 3 MGy.[64]

Figure 4.4 also highlights another issue connected with quoting doses. For the Gaussian beams there are several possible options, including the maximum dose, the average dose for the whole crystal and the average dose of the irradiated volume. To provide a more robust dose value for comparison of different experiments, a new dose metric, Diffraction Weighted Dose (DWD), is output by RADDOSE-3D. DWD is a fluence-weighted dose, which combines information from the aggregate value of the dose up to a


X-rays
X-rays
X-rays
Dose (MGy)
6
5
4
3
2
1
0
(a)
(b)
(c)
(d)
(e)
(f)
10⁹ ph/s
10⁸ ph/s
10⁵ ph/s
x
y

given time within each volume element (voxel) of the crystal with the way the crystal is being irradiated at that moment. In a validation experiment, very similar intensity decay rates for three different sized beams were obtained when I_n/I_1, was plotted against DWD.[62]

Because beam profiles have such a great influence on dose values, RADDOSE-3D has been modified so that experimentally determined profiles can now be read in order to increase the accuracy of the calculation for particular measurements. Other recent developments are that crystals can now be described as polygons rather than just spheres or cuboids, and the program can read in a PDB file to extract the amino acid content if the structure has been deposited.

RADDOSE-3D, as did RADDOSE v3, now accounts for the dose absorbed through Compton electron production, so it can be used to model experiments at high incident beam energies However, a remaining issue with the code is that unlike RADDOSE v2, it does not account for the possible escape of fluorescent X-rays emitted by any heavy atoms contained in the crystal. Another aspect which has been very lately included in the code is that it now takes into account the possible escape of the primary photoelectron from the crystal. It has previously been pointed out that for very small crystals, there was a probability that the photoelectron could have a longer range than the crystal size and so could escape, and this would reduce the dose, allowing small crystals to have a longer lifetime in the beam.[65] Currently, for the first time, the latest version of RADDOSE-3D (soon to be released) contains an analytical model for this phenomenon that is robust for incident beam energies up to 20 keV. Simulations using it show that at a beam energy of 12.4 keV, for crystals larger than $(20\ \mu m)^3$, the dose reduction is only 3% whereas the dose is halved for a $(5\ \mu m)^3$ crystal. As the incident beam energy is increased, the probability of photoelectron escape rises since they carry more energy. However, this will not be true for a crystal held within a

Figure 4.4 The effect of varying the incident beam profile on the dose distribution for an irregular polyhedron shaped crystal formed by distorting a cube of dimension $(100\ \mu m)^3$ leaving the total volume the same at $10^6\ \mu m^3$. In (a)–(c) the distribution of dose throughout the crystal has been calculated by RADDOSE-3D[57] and displayed using the open source Paraview software package (https://www.paraview.org). In (a)–(b) the beam has been modelled as Gaussian shaped, and the FWHM has been varied as: (a) 20 μm×20 μm and (b) 60 μm×60 μm. In (c) a uniform top-hat shaped beam distribution has been modelled. In (d)–(f) the corresponding two-dimensional beam profiles for each exposure strategy have been illustrated. In each plot, $(x,y) = (0,0)$ corresponds to the intersection of the X-ray beam direction and the single rotation axis. All other beam parameters have been kept constant (energy: 12.4 keV, flux 5×10^{11} photons s^{-1}, rectangular collimation: 100 μm×100 μm). In all simulations, the crystal has been exposed for 100 s throughout a 360° rotation about the *y*-axis (shown in each plot). The direction of incident beam with respect to the initial orientation of the crystal has been shown (yellow arrow).

cryoloop completely surrounded by vitrified cryobuffer which is also in the beam path, since approximately as many photoelectrons will enter the crystal from the buffer as will escape from it. The RADDOSE-3D photoelectron escape model that has been incorporated into the version soon to be released is only valid when the effects of the surrounding buffer are assumed to be negligible.[66]

4.5 X-ray Induced Changes in Chromophore-containing Proteins at 100 K

Proteins containing a chromophore, such as a conjugated π-electron system or a metal complex,[67] are particularly prone to suffer from specific radiation damage. Chromophore-specific damage can be identified and followed by *in crystallo* spectroscopic techniques such as X-ray absorption spectroscopy,[68,69] optical absorption[34] and Raman spectroscopy[70] and electron paramagnetic resonance.[71] Radiation-induced chromophore damage already occurs at absorbed doses up to three orders of magnitude lower[68] than the 30 MGy limit for specific and global damage at 100 K.[27] The dose at which 90% of the unreduced state remains has been baptized *spectroscopic life dose*.[72] The spectroscopic life dose at 100 K is 0.01 MGy for the ferric (oxidized) haem iron in metmyoglobin[72] and 0.04 MGy for retinal in the membrane protein bacteriorhodopsin.[73] Other examples of specific chromophore damage are haem reduction in high-molecular weight cytochrome c,[74] in a photosynthetic reaction center[75] and in cytochrome c peroxidase,[76] redox changes in a methylamine dehydrogenase,[77] photobleaching of a fluorescent protein[44] and deprotonation of the bilin chromophore in a phytochrome.[78]

Strategies have been developed to minimize radiation-induced changes to chromophores in conventional MX experiments, such as composite[79] and multi-crystal data collection schemes,[80] adding scavengers[81] or lowering the sample temperature to 40 K, where metal reduction is reduced 30-fold compared to 110 K.[82] Alternatively, serial crystallography at XFELs (see Chapter 2) or synchrotron sources (see Chapter 1) can be used to spread out the dose over tens of thousands of crystals.

4.6 Global and Specific Radiation Damage Above 100 K and at Room Temperature

The success of cryo-cooling[2] in mitigating radiation damage is due to the diffusion of radicals such as $OH^{\bullet}$ or more bulky species being abolished at 100 K (see above). Indeed, during flash-cooling, the crystal solvent forms an amorphous solid and internal macromolecular motions are quenched as they undergo a glass transition and a dynamical transition, respectively.[83–85] These transitions occur at around 200 K, so that radiation-damage proceeds at low rates below and at high rates above that temperature.[86]

The (cryo-) temperature-dependence of radical diffusion can be exploited to trigger a reaction within a crystalline macromolecule by X-ray irradiation that is then followed by temperature-controlled crystallography.[87] Such a data-collection strategy has been applied to study the reaction pathway of P450cam cytochrome.[88] The reaction was triggered by radiolytically generated electrons and its progression controlled by transiently thawing the crystals to room temperature in order to unlock the macromolecular motions. In a second example, a non-hydrolysable substrate-analogue in complex with crystalline acetylcholinesterase was radiolysed at 100 and 155 K. Only at 155 K was the macromolecular flexibility large enough for products to escape the active site and for associated conformational changes to occur.[89]

Since the crystal lifetime at RT is reduced by up to two orders of magnitude compared to that at 100 K,[9,90,91] it is difficult, if not impossible, to follow specific radiation damage as a function of absorbed dose by collecting a series of consecutive RT datasets on the same part of a single crystal as is traditionally done at 100 K. Nevertheless, clear evidence for specific damage to disulfide bonds at RT has been presented.[90] The recent adaptation of room-temperature serial crystallography developed at XFELs (see Chapter 2) for data collection at synchrotron sources (see Chapter 1) now allows the dose to be spread over many crystals and has opened the door to tracking specific damage at RT. The first such studies provided conflicting results, with specific damage being clearly[92] or barely visible[93] or absent.[55] More RT serial synchrotron crystallography experiments will have to be carried out to address the issue of specific damage and to verify earlier observations that suggest global radiation damage can be partially outrun by collecting data at high dose rates.[90,94,95]

4.7 Recruitment of Radiation-induced Changes to Study Macromolecular Function

Although radiation damage is usually deleterious to MX structure solution (but see 4.8), it can actually be utilized to elucidate the structural intermediates in reaction pathways. This potential is exemplified by studies of P450cam cytochrome described in the previous section.[88] Another trailblazing such study was that involving horse radish peroxidase, where redox intermediates formed as a result of beam-induced electrons. A carefully designed data collection protocol allowed a dose series to be assembled from 10 different datasets taken at different regularly spaced starting angles and then sections of them were assembled into composite 'same dose' complete datasets for structure solution.[79] In a more recent study, serial synchrotron crystallography (see Chapter 1) at 100 K was utilized by Horrell and coworkers to track enzyme catalysis in copper nitrite reductase.[96]

As already mentioned in Section 4.5, such studies can benefit greatly from the use of complementary techniques to monitor the enzymatic or redox state of the protein, either simultaneously during X-ray irradiation (such as

on-line microspectrophotometry[34]) or by characterizing the crystal preceding and post the diffraction data collection. Through MX in conjunction with *in crystallo* Raman spectroscopy and QM/MM simulations, the catalytic cycle of the non-chromophore-containing urate oxidase, again initiated following X-ray absorption, has been elucidated.[97]

Very recently a study tracking the intermediates in laccase, a multicopper oxidase, was reported. Again, using well diffracting crystals (1.35 Å), electrons induced by the X-ray beam triggered the reduction of molecular oxygen to water, which was tracked with increasing absorbed dose. Although the resulting structures in the dose series are mixtures of different states of the enzyme–substrate complex, these could be deconvoluted to shed light on the reduction mechanism.[98] Structural details of haem–ligand intermediate states in protein crystals have long been of great interest, and the haem state can conveniently be monitored using online microspectrophotometry synergistically in concert with MX during its reduction by the X-ray beam. An oxyferrous complex of the dye-decolorizing-type peroxidase A (DtpA) has been structurally characterised using this combination. Also studied was a five-coordinate haem protein cytochrome c0 for which both MX and single-crystal resonance Raman (SCRR) data were recorded to identify the intermediate states, following photoreduction by a 405 nm laser.[99]

4.8 Radiation-damage Induced Phasing

The effects of radiation damage in MX are not solely deleterious to structure solution. It has been elegantly shown by Ravelli and coworkers[100] that the specific structural damage inflicted by the X-ray beam can be used for protein structure solution by 'Radiation-damage Induced Phasing' (RIP). In this method carried out at 100 K, a 'low dose' relatively undamaged dataset is first collected, and the crystal is then subjected to an 'X-ray burn' up to a dose of several MGy which breaks the disulfide bonds, disordering their sulfur atoms. Another low dose dataset is then collected and the intensity differences from the first dataset used to locate the 'heavy' (sulfur) atoms that are now disordered, in a manner similar to a single isomorphous replacement phasing procedure in reverse (destroying the intrinsic heavy atom signal rather than adding a heavy atom to the crystal). RIP can also be utilized in concert with SAD,[101] and this has resulted in the structure solution of the $^{2}F^{3}F$ domains of human fibronectin,[102] a particularly challenging case which had previously been intransigent to any other phasing method.[103] The specific damage inflicted by UV light as opposed to X-rays can also be used for the 'burn' stage of RIP (UV-RIP).[104]

Although experimentally the RIP method can be challenging (the 'burn' must not inflict too much global damage, thus doses must be carefully estimated and controlled), and well diffracting crystals are required, it has been exploited to good effect through using the radiation induced breakage of an endogenous C–I bond to solve the racemic structure of the antimicrobial protein snakin-1, isolated from potatoes.[105]

4.9 Does Radiation Damage Depend on Dose Rate and/or on the Incident Beam Energy at 100 K?

New 4th generation synchrotron sources are now starting to come on line, and they will bring the issue of radiation damage in MX into even sharper focus. Their flux densities will be more than an order of magnitude greater than those currently available, such that the 100 K experimental dose limit of 30 MGy will be reached in around 10 ms. One of the important unknown issues for MX at these new sources is whether there will be dose rate effects for crystals irradiated at 100 K, and whether sample heating will be significant.

Regarding the dose rate question, in 1994 at a 2nd generation source, Gonzalez and Nave obtained similar intensity decay rates at 100 K using two different beam attenuations.[10] At 3rd generation X-ray sources, the available evidence points to there being no significant global damage dose rate effects when using flux densities up to 10^{15} photons s^{-1} mm^{-2}.[106] However, Owen and coworkers[27] noted a small effect (a 10% decrease in $D_{1/2}$ when the dose rate was increased by a factor of 10). In another study, no difference in intensity decay was observed at dose rates ranging by almost two orders of magnitude, but analysis of the specific damage seen in electron density maps suggested that the damage was more severe at the higher dose rates, and it was hypothesized that this might be due to differences in the steady-state radical concentration.[107] However, as mentioned in Section 4.6, at RT and close to it, there have been observations of significant dose rate effects.

The second issue concerns the possible effects of crystal heating in higher flux density beams. Detailed fluid-dynamics modelling of the heat transfer processes for crystals held at 100 K and irradiated at 3rd generation synchrotrons showed that beam induced sample heating would not be significant.[108] However, this issue may become pertinent at 4th generation MX beamlines, since if the crystal is heated to above 110 K by the beam, hydroxyl radicals will be able to move and start to diffuse through the sample spreading damage.

Another frequently posed question in relation to radiation damage in MX is whether or not its rate depends on the energy, E_{inc}, of the incident beam, and as yet a general consensus on this has not emerged. As highlighted by Arndt,[109] the cross-section for diffraction is higher at lower E_{inc} (corresponding to higher wavelengths, λ, since $E_{inc} = hc/\lambda$ where h is Planck's constant and c is the speed of light; this simplifies to $E_{inc}[\text{keV}] = 12.4/\lambda[\text{Å}]$). However, the probability of photoelectric absorption also increases at low E_{inc}, and the ratio of diffraction to absorption events remains almost constant over the range of E_{inc} normally used in MX ($E_{inc} \sim 7$ to 14 keV, $\lambda \sim 1.7$ to 0.9 Å). Theoretically, there should be no change in the damage rates at different E_{inc}, despite much anecdotal evidence from RT data collections to the contrary. This analysis was experimentally tested by Weiss and coworkers using elastase crystals at 100 K at $E_{inc} = 12.4$ and 6.2 keV, with an analysis of

the disulfide bonds and cadmium ions in the resulting structures showing no difference in damage rates.[110] Another study by Shimizu and coworkers used CEWL crystals and monitored their global and specific damage behaviour at six different E_{inc} values between 6.5 and 33.0 keV, finding no difference in any of the reciprocal or real space metrics, including the atomic *B*-factors of the refined structures.[111] However, in an experiment on CEWL crystals at $E_{inc} = 9$ and 14 keV, it was found by Homer and coworkers that the electron density of cysteines in the disulfide bonds (measured in electrons $Å^{-3}$ MGy^{-1}) decreased faster at 14 keV than at 9 keV, although no difference was noted in the rate of methionine damage.[112]

Contrary to the results described above which found no significant global damage variation with E_{inc}, Fourme and coworkers stated that their data collection efficiency was eight times greater at 33 keV than at 8 keV,[113] and Liebschner and coworkers also reported lower (but similar) decay rates at 12.66 and 19.00 keV compared with 6.33 keV for repeated 2° wedges of thaumatin data ($D_{1/2}$ ~11 MGy and 7.5 MGy respectively, at 2.45 Å resolution).[114]

From this brief summary, it is clear that it is not yet possible to draw robust conclusions concerning the influence of E_{inc} on the rates of radiation damage in MX.

4.10 How Can Radiation Damage Be Minimised?

The question asked by many researchers is how radiation damage can be minimized in their experiments. There are various strategies and some general guidelines that can be tried, some of which have been systematically tested. An initial consideration is the priority for the particular data collection, and this will depend on the aim of the experiment; is it to obtain a complete dataset or the highest possible resolution from the crystals? If a large supply of crystals is available, there are now good protocols for composite data collection and for merging data from many samples together.[115] In this case a few high-resolution images can be obtained from each crystal before they are too badly radiation damaged. As discussed above, at 100 K there is not thought to be a significant dose rate effect, so the radiation damage inflicted by taking, for instance, five short exposure images should be the same as taking one image for five times longer in order to obtain the highest resolution possible. Serial synchrotron crystallography, where one still image is collected per crystal, is also now gaining momentum as a technique (see Chapter 1) and strategies to minimize radiation damage are being investigated.[116]

Another worthwhile strategy is to consider other dose spreading data collection protocols such as the helical scans (for rod shaped crystals) mentioned above. If available, it is advantageous to irradiate the crystal as homogeneously as possible with a top-hat shaped beam rather than one with a Gaussian profile (the latter can be defocused and slits used to select the top of the Gaussian to allow more even irradiation of the sample). Matching the beam size to the crystal size will reduce the background scatter so that

high-resolution reflections will be easier to detect even as the crystal becomes damaged.

Critically, the absorption coefficient of the crystal can be reduced by back soaking any non-ordered heavy atoms present in the buffer, or when soaking in heavy atom compounds if phasing is being attempted by multi-isomorphous replacement methods. Back-soaking can be performed prior to cryo-cooling. If the absorption coefficient can be reduced in this way, the dose is concomitantly lower, and the crystal will last longer in the beam but will give the same intensity of ordered diffraction as before the back-soaking. For instance, there is arsenic in cacodylate buffer which has a significantly bigger X-ray cross-section than that of organic buffers. To quantify the effect of buffers on dose, RADDOSE was used by Holton[117] to calculate the 'dose doubling' concentrations of commonly used MX crystallisation buffers, which is only 350 mM for arsenic and as little as 100 mM for gold and platinum.

Finally, crystals can be co-crystallised or soaked in scavengers before being cryo-cooled, in order to reduce the level of radicals (particularly electrons and $OH^{\bullet}$ moieties) present in the solvent channels. While there is some anecdotal evidence for the efficacy of this strategy, systematic studies have failed to reach a consensus, and different researchers have reported conflicting results for tests with the same scavenger on the same protein crystal type. Holton[117] has advocated a criterion for judging the results of experiments designed to reduce rates of radiation damage: unless there is an improvement of at least a factor of two, the results are probably not significant. Very few scavengers, even when deemed to prolong crystal lifetime, meet this criterion (*i.e.* they do not increase $D_{1/2}$ by greater than twofold). The one exception is 1,4-benzoquinone, which has been reported to increase CEWL crystal $D_{1/2}$ values at RT by a factor of nine[118] and concomitantly significantly reduce the specific damage, particularly at disulfide bonds. Of nineteen putative scavengers tested by Kmetko *et al.*[119] at 100 K, none were found to be effective, and at RT six sensitised the crystals, 12 were ineffective, and only the very efficient electron scavenger, $NaNO_3$, was judged to be effective. This is in contrast to work by De la Mora *et al.*, who found that $NaNO_3$ at 100 K almost doubled $D_{1/2}$ and protected disufide bonds so that they were undamaged at 5 times the dose at which damage was evident for the native unsoaked crystal. In this case the electron density maps showed that one of the disulfide bonds (Cys6–Cys127) had a bound nitrate anion next to it (see Figure 4.5), which was seen to be sequentially reduced to NO_2 and then NO as the dose increased. The nitrate is very efficiently scavenging the mobile secondary electrons which can quantum-mechanically tunnel around the protein structure at 100 K and is intercepting them before they reach the highly electron-affinic disulfide bonds. The electron density maps in Figure 4.5 give snapshots of this radiation chemistry in action.

Since the comprehensive summary of all MX scavenger experiments published before 2013 was compiled by Allan and coworkers,[120] to our knowledge there has been only one further reported study. This tested the efficacy of uridine at various concentrations and at RT it was deemed

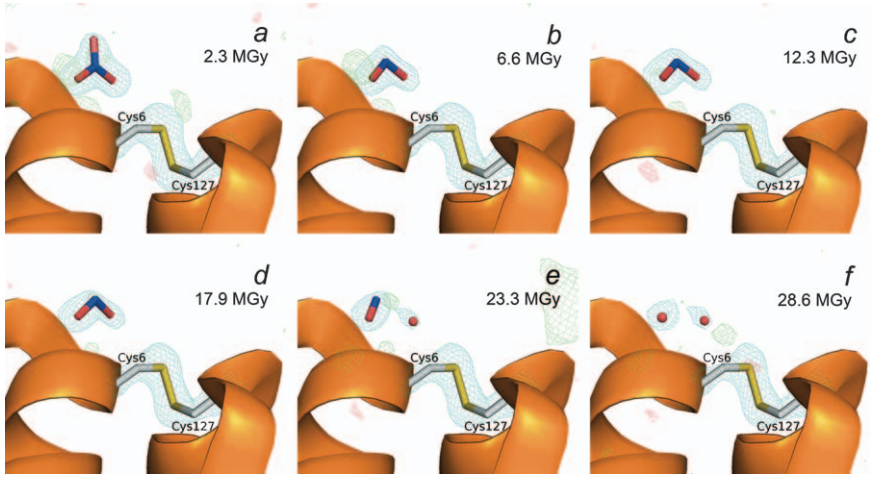


Figure 4.5 Radiation chemistry in action. Series of structures at increasing dose derived from data collected at 100 K from a CEWL crystal soaked in 0.5 M $NaNO_3$ for 4 min before cryo-cooling. Panels (a)–(f) show structures refined from consecutive datasets 1–6 at 2.3 to 28.6 MGy, respectively, for the Cys6–Cys127 disulfide bond. The 2|Fo|−|Fc| electron density map (light blue) is contoured at 1.0 σ and the |Fo|−|Fc| electron density map is contoured at ±3.0 σ (green, positive; red, negative). Mobile electrons produced by the X-ray beam appear to reduce a bound nitrate anion to NO_2 (6.6 MGy) and then to NO (23.3 MGy). The disulfide bond is protected from damage until the nitrate scavenging capacity is exhausted. The PDB codes of the refined models are 2BYH, 2BYI, 2BYJ, 2BYL, 2BYM, 2BYN, and are shown in panels (a) to (f), respectively.[30]

effective in suppressing global effects, although the maximum extension of life dose (at 1 M) was only 70% (*i.e.* factor of 1.7). However, uridine did not protect against specific damage to susceptible residues (Cys, Asp, Glu). At 100 K, there were no protective effects detected at all.[121]

In conclusion on scavengers, they are not in routine use but may be worth trying for particularly difficult cases. However, they are not guaranteed to give improvement in crystal life-times. Of course, many buffer components used to crystallise proteins have high rate constants for scavenging hydroxyl radicals, especially the soluble organics such as PEG, glycerol and ethylene glycol. To provide noticeable improvement in dose tolerance, any added scavenger would have to be both highly efficient and also present in sufficient quantity to outcompete these inbuilt agents.

4.11 Radiation Damage in Serial Femtosecond Crystallography at XFELs

Serial femtosecond crystallography[15,16] exploits the short and brilliant femtosecond pulses provided by XFELs to produce diffraction data before

chemical damage has the time to develop in the sample, so-called diffraction-before-destruction.[14] The site-specific damage often seen in protein structures solved from synchrotron data collected at 100 K can indeed be avoided by SFX,[122] resulting in radiation-damage-free structural models.[123,124] However, at very high absorbed doses (*i.e.* close to 3 GGy[125]) global radiation damage occurs[125,126] and evidence for specific damage has also been presented.[125] Such specific damage at high doses is manifested for example by reduced electron density for the iron atoms in [4Fe – 4S] clusters of crystalline ferredoxin[127] and by damage at sulfur sites in crystalline cathepsin B revealed by a difference Fourier map based on data collected at high and low X-ray fluence.[128] The risk of generating global and specific damage can be reduced by collecting SFX data with short and attenuated XFEL pulses.

Acknowledgements

We thank Charles Bury and Eugenio De la Mora for drawing the figures, Markus Gerstel for mining the PDB for the temperature of data collection information, and Ian Carmichael for his constructive comments on this Chapter. We gratefully acknowledge the contribution of the ESRF to the ongoing effort to understand the systematics of radiation damage in MX, since we have been granted regular beam time since 2000 under the auspices of the Radiation Damage BAG.

References

1. T. L. Blundell and L. N. Johnson, *Protein Crystallography*, Academic Press, 1976.
2. H. Hope, *Acta Crystallogr., Sect. B*, 1988, **44**, 22–26.
3. T. Teng, *J. Appl. Crystallogr.*, 1990, **23**, 387–391.
4. D. W. Rodgers, *Structure*, 1994, **2**, 1135–1140.
5. D. W. Rodgers, *Methods in Enzymology*, Academic Press, 1997, vol. 276, pp. 183–203.
6. E. F. Garman and T. R. Schneider, *J. Appl. Crystallogr.*, 1997, **30**, 211–237.
7. E. Garman, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 1999, **55**, 1641–1653.
8. J. Cosier and A. M. Glazer, *J. Appl. Crystallogr.*, 1986, **19**, 105–107.
9. C. Nave and E. F. Garman, *J. Synchrotron Radiat.*, 2005, **12**, 257–260.
10. A. Gonzalez and C. Nave, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 1994, **50**, 874–877.
11. M. Weik, R. B. Ravelli, G. Kryger, S. McSweeney, M. L. Raves, M. Harel, P. Gros, I. Silman, J. Kroon and J. L. Sussman, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 623–628.
12. R. B. Ravelli and S. M. McSweeney, *Struct. Folding Des.*, 2000, **8**, 315–328.

13. W. P. Burmeister, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2000, **56**, 328–341.
14. R. Neutze, R. Wouts, D. van der Spoel, E. Weckert and J. Hajdu, *Nature*, 2000, **406**, 752–757.
15. H. N. Chapman, P. Fromme, A. Barty, T. A. White, R. A. Kirian, A. Aquila, M. S. Hunter, J. Schulz, D. P. DePonte, U. Weierstall, R. B. Doak, F. R. Maia, A. V. Martin, I. Schlichting, L. Lomb, N. Coppola, R. L. Shoeman, S. W. Epp, R. Hartmann, D. Rolles, A. Rudenko, L. Foucar, N. Kimmel, G. Weidenspointner, P. Holl, M. Liang, M. Barthelmess, C. Caleman, S. Boutet, M. J. Bogan, J. Krzywinski, C. Bostedt, S. Bajt, L. Gumprecht, B. Rudek, B. Erk, C. Schmidt, A. Homke, C. Reich, D. Pietschner, L. Struder, G. Hauser, H. Gorke, J. Ullrich, S. Herrmann, G. Schaller, F. Schopper, H. Soltau, K. U. Kuhnel, M. Messerschmidt, J. D. Bozek, S. P. Hau-Riege, M. Frank, C. Y. Hampton, R. G. Sierra, D. Starodub, G. J. Williams, J. Hajdu, N. Timneanu, M. M. Seibert, J. Andreasson, A. Rocker, O. Jonsson, M. Svenda, S. Stern, K. Nass, R. Andritschke, C. D. Schroter, F. Krasniqi, M. Bott, K. E. Schmidt, X. Wang, I. Grotjohann, J. M. Holton, T. R. Barends, R. Neutze, S. Marchesini, R. Fromme, S. Schorb, D. Rupp, M. Adolph, T. Gorkhover, I. Andersson, H. Hirsemann, G. Potdevin, H. Graafsma, B. Nilsson and J. C. Spence, *Nature*, 2011, **470**, 73–77.
16. I. Schlichting, *IUCrJ*, 2015, **2**, 246–255.
17. C. Nave, *Radiat. Phys. Chem.*, 1995, **45**, 483–490.
18. R. Sanishvili, D. W. Yoder, S. B. Pothineni, G. Rosenbaum, S. Xu, S. Vogt, S. Stepanov, O. A. Makarov, S. Corcoran, R. Benn, V. Nagarajan, J. L. Smith and R. F. Fischetti, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 6127–6132.
19. P. O'Neill, D. L. Stevens and E. F. Garman, *J. Synchrotron Radiat.*, 2002, **9**, 329–332.
20. E. Garman, *Acta Crystallogr., Sect. D*, 2010, **66**, 339–351.
21. R. L. Owen, D. Axford, J. E. Nettleship, R. J. Owens, J. I. Robinson, A. W. Morgan, A. S. Dore, G. Lebon, C. G. Tate, E. E. Fry, J. Ren, D. I. Stuart and G. Evans, *Acta Crystallogr., Sect. D*, 2012, **68**, 810–818.
22. G. D. Jones, J. S. Lea, M. C. Symons and F. A. Taiwo, *Nature*, 1987, **330**, 772–773.
23. T. Teng and K. Moffat, *J. Synchrotron Radiat.*, 2000, **7**, 313–317.
24. J. Murray and E. Garman, *J. Synchrotron Radiat.*, 2002, **9**, 347–354.
25. R. B. Ravelli, P. Theveneau, S. McSweeney and M. Caffrey, *J. Synchrotron Radiat.*, 2002, **9**, 355–360.
26. A. Meents, S. Gutmann, A. Wagner and C. Schulze-Briese, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 1094–1099.
27. R. L. Owen, E. Rudiño-Piñera and E. F. Garman, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 4912–4917.
28. J. Kmetko, N. S. Husseini, M. Naides, Y. Kalinin and R. E. Thorne, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2006, **62**, 1030–1038.

29. K. Diederichs, S. McSweeney and R. B. Ravelli, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2003, **59**, 903–909.
30. E. De la Mora, I. Carmichael and E. F. Garman, *J. Synchrotron Radiat.*, 2011, **18**, 346–357.
31. J. R. Helliwell, *J. Cryst. Growth*, 1988, **90**, 259–272.
32. K. A. Sutton, P. J. Black, K. R. Mercer, E. F. Garman, R. L. Owen, E. H. Snell and W. A. Bernhard, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2013, **69**, 2381–2394.
33. M. Weik, J. Berges, M. L. Raves, P. Gros, S. McSweeney, I. Silman, J. L. Sussman, C. Houee-Levin and R. B. Ravelli, *J. Synchrotron Radiat.*, 2002, **9**, 342–346.
34. J. McGeehan, R. B. Ravelli, J. W. Murray, R. L. Owen, F. Cipriani, S. McSweeney, M. Weik and E. F. Garman, *J. Synchrotron Radiat.*, 2009, **16**, 163–172.
35. P. Wisniowski, I. Carmichael, R. W. Fessenden and G. L. Hug, *J. Phys. Chem. A*, 2002, **106**, 4573–4580.
36. E. Fioravanti, F. M. Vellieux, P. Amara, D. Madern and M. Weik, *J. Synchrotron Radiat.*, 2007, **14**, 84–91.
37. M. Gerstel, C. M. Deane and E. F. Garman, *J. Synchrotron Radiat.*, 2015, **22**, 201–212.
38. C. S. Bury, I. Carmichael and E. F. Garman, *J. Synchrotron Radiat.*, 2017, **24**, 7–18.
39. A. P. Dubnovitsky, R. B. Ravelli, A. N. Popov and A. C. Papageorgiou, *Protein Sci.*, 2005, **14**, 1498–1507.
40. Y. Matsui, K. Sakai, M. Murakami, Y. Shiro, S. Adachi, H. Okumura and T. Kouyama, *J. Mol. Biol.*, 2002, **324**, 469–481.
41. R. Kort, K. J. Hellingwerf and R. B. Ravelli, *J. Biol. Chem.*, 2004, **279**, 26417–26424.
42. A. Mees, T. Klar, P. Gnau, U. Hennecke, A. P. Eker, T. Carell and L. O. Essen, *Science*, 2004, **306**, 1789–1793.
43. B. Sjoblom, M. Polentarutti and K. Djinovic-Carugo, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 10609–10613.
44. V. Adam, P. Carpentier, S. Violot, M. Lelimousin, C. Darnault, G. U. Nienhaus and D. Bourgeois, *J. Am. Chem. Soc.*, 2009, **131**, 18063–18065.
45. K. Takeda, Y. Matsui, N. Kamiya, S. Adachi, H. Okumura and T. Kouyama, *J. Mol. Biol.*, 2004, **341**, 1023–1037.
46. C. Wickstrand, R. Dods, A. Royant and R. Neutze, *Biochim. Biophys. Acta, Gen. Subj.*, 2015, **1850**, 536–553.
47. C. Bury, E. F. Garman, H. M. Ginn, R. B. G. Ravelli, I. Carmichael, G. Kneale and J. E. McGeehan, *J. Synchrotron Radiat.*, 2015, **22**, 213–224.
48. C. S. Bury, J. E. McGeehan, A. A. Antson, I. Carmichael, M. Gerstel, M. B. Shevtsov and E. F. Garman, *Acta Crystallogr., Sect. D*, 2016, **72**, 648–657.
49. U. Chinte, B. Shah, Y. S. Chen, A. A. Pinkerton, C. A. Schall and B. L. Hanson, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2007, **63**, 486–492.

50. A. Meents, A. Wagner, R. Schneider, C. Pradervand, E. Pohl and C. Schulze-Briese, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2007, **63**, 302–309.
51. R. L. Owen, J. M. Holton, C. Schulze-Briese and E. F. Garman, *J. Synchrotron Radiat.*, 2009, **16**, 143–151.
52. R. Henderson, *Proc. R. Soc. London, Ser. B*, 1990, **241**, 6–8.
53. E. F. Garman and G. W. Grime, *Prog. Biophys. Mol. Biol.*, 2005, **89**, 173–205.
54. M. R. Howells, T. Beetz, H. N. Chapman, C. Cui, J. M. Holton, C. J. Jacobsen, J. Kirz, E. Lima, S. Marchesini, H. Miao, D. Sayre, D. A. Shapiro, J. C. Spence and D. Starodub, *J. Electron Spectrosc. Relat. Phenom.*, 2009, **170**, 4–12.
55. P. Roedig, R. Duman, J. Sanchez-Weatherby, I. Vartiainen, A. Burkhardt, M. Warmer, C. David, A. Wagner and A. Meents, *J. Appl. Crystallogr.*, 2016, **49**, 968–975.
56. J. W. Murray, E. Rudiño-Piñera, R. L. Owen, M. Grininger, R. B. Ravelli and E. F. Garman, *J. Synchrotron Radiat.*, 2005, **12**, 268–275.
57. O. B. Zeldin, M. Gerstel and E. F. Garman, *J. Appl. Crystallogr.*, 2013, **46**, 1225–1230.
58. K. S. Paithankar, R. L. Owen and E. F. Garman, *J. Synchrotron Radiat.*, 2009, **16**, 152–162.
59. K. S. Paithankar and E. F. Garman, *Acta Crystallogr., Sect. D*, 2010, **66**, 381–388.
60. O. B. Zeldin, M. Gerstel and E. F. Garman, *J. Synchrotron Radiat.*, 2013, **20**, 49–57.
61. D. Flot, T. Mairs, T. Giraud, M. Guijarro, M. Lesourd, V. Rey, D. van Brussel, C. Morawe, C. Borel, O. Hignette, J. Chavanne, D. Nurizzo, S. McSweeney and E. Mitchell, *J. Synchrotron Radiat.*, 2010, **17**, 107–118.
62. O. B. Zeldin, S. Brockhauser, J. Bremridge, J. M. Holton and E. F. Garman, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 20551–20556.
63. T. Krojer and F. von Delft, *J. Synchrotron Radiat.*, 2011, **18**, 387–397.
64. S. L. Storm, F. Dall'Antonia, G. Bourenkov and T. R. Schneider, *J. Synchrotron Radiat.*, 2017, **24**, 19–28.
65. C. Nave and M. A. Hill, *J. Synchrotron Radiat.*, 2005, **12**, 299–303.
66. C. S. Bury, J. C. Brooks-Bartlett, S. P. Walsh and E. F. Garman, *Protein Sci.*, 2018, **27**, 217–228.
67. T. Beitlich, K. Kuhnel, C. Schulze-Briese, R. L. Shoeman and I. Schlichting, *J. Synchrotron Radiat.*, 2007, **14**, 11–23.
68. J. Yano, J. Kern, K. D. Irrgang, M. J. Latimer, U. Bergmann, P. Glatzel, Y. Pushkar, J. Biesiadka, B. Loll, K. Sauer, J. Messinger, A. Zouni and V. K. Yachandra, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 12047–12052.
69. J. M. Holton, *J. Synchrotron Radiat.*, 2007, **14**, 51–72.
70. P. Carpentier, A. Royant, J. Ohana and D. Bourgeois, *J. Appl. Crystallogr.*, 2007, **40**, 1113–1122.
71. L. M. Utschig, S. D. Chemerisov, D. M. Tiede and O. G. Poluektov, *Biochemistry*, 2008, **47**, 9251–9257.

72. H. P. Hersleth and K. K. Andersson, *Biochim. Biophys. Acta*, 2011, **1814**, 785–796.
73. V. Borshchevskiy, E. Round, I. Erofeev, M. Weik, A. Ishchenko, I. Gushchin, A. Mishin, D. Willbold, G. Buldt and V. Gordeliy, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2014, **70**, 2675–2685.
74. M. Sato, N. Shibata, Y. Morimoto, Y. Takayama, K. Ozawa, H. Akutsu, Y. Higuchi and N. Yasuoka, *J. Synchrotron Radiat.*, 2004, **11**, 113–116.
75. R. H. Baxter, B. L. Seagle, N. Ponomarenko and J. R. Norris, *J. Am. Chem. Soc.*, 2004, **126**, 16728–16729.
76. A. Echalier, C. F. Goodhew, G. W. Pettigrew and V. Fulop, *Structure*, 2006, **14**, 107–117.
77. A. R. Pearson, R. Pahl, E. G. Kovaleva, V. L. Davidson and C. M. Wilmot, *J. Synchrotron Radiat.*, 2007, **14**, 92–98.
78. F. Li, E. S. Burgie, T. Yu, A. Heroux, G. C. Schatz, R. D. Vierstra and A. M. Orville, *J. Am. Chem. Soc.*, 2015, **137**, 2792–2795.
79. G. I. Berglund, G. H. Carlsson, A. T. Smith, H. Szoke, A. Henriksen and J. Hajdu, *Nature*, 2002, **417**, 463–468.
80. H. Aoyama, K. Muramoto, K. Shinzawa-Itoh, K. Hirata, E. Yamashita, T. Tsukihara, T. Ogura and S. Yoshikawa, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 2165–2169.
81. S. Macedo, M. Pechlaner, W. Schmid, M. Weik, K. Sato, C. Dennison and K. Djinovic-Carugo, *J. Synchrotron Radiat.*, 2009, **16**, 191–204.
82. M. C. Corbett, M. J. Latimer, T. L. Poulos, I. F. Sevrioukova, K. O. Hodgson and B. Hedman, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2007, **63**, 951–960.
83. D. Vitkup, D. Ringe, G. A. Petsko and M. Karplus, *Nat. Struct. Biol.*, 2000, **7**, 34–38.
84. M. Weik, G. Kryger, A. M. Schreurs, B. Bouma, I. Silman, J. L. Sussman, P. Gros and J. Kroon, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2001, **57**, 566–573.
85. G. Schiro, Y. Fichou, F. X. Gallat, K. Wood, F. Gabel, M. Moulin, M. Hartlein, M. Heyden, J. P. Colletier, A. Orecchini, A. Paciaroni, J. Wuttke, D. J. Tobias and M. Weik, *Nat. Commun.*, 2015, **6**, 6490.
86. M. Warkentin and R. E. Thorne, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2010, **66**, 1092–1100.
87. M. Weik and J.-P. Colletier, *Acta Crystallogr., Sect. D*, 2010, **66**, 437–446.
88. I. Schlichting, J. Berendzen, K. Chu, A. M. Stock, S. A. Maves, D. E. Benson, R. M. Sweet, D. Ringe, G. A. Petsko and S. G. Sligar, *Science*, 2000, **287**, 1615–1622.
89. J. P. Colletier, D. Bourgeois, B. Sanson, D. Fournier, J. L. Sussman, I. Silman and M. Weik, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 11742–11747.
90. R. J. Southworth-Davies, M. A. Medina, I. Carmichael and E. F. Garman, *Structure*, 2007, **15**, 1531–1541.
91. M. Warkentin, J. B. Hopkins, R. Badeau, A. M. Mulichak, L. J. Keefe and R. E. Thorne, *J. Synchrotron Radiat.*, 2013, **20**, 7–13.

92. N. Coquelle, A. S. Brewster, U. Kapp, A. Shilova, B. Weinhausen, M. Burghammer and J.-P. Colletier, *Acta Crystallogr., Sect. D*, 2015, **71**, 1184–1196.
93. S. Russi, A. Gonzalez, L. R. Kenner, D. A. Keedy, J. S. Fraser and H. Van den Bedem, *J. Synchrotron Radiat.*, 2017, **24**, 73–82.
94. M. Warkentin, R. Badeau, J. B. Hopkins, A. M. Mulichak, L. J. Keefe and R. E. Thorne, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2012, **68**, 124–133.
95. R. L. Owen, N. Paterson, D. Axford, J. Aishima, C. Schulze-Briese, J. Ren, E. E. Fry, D. I. Stuart and G. Evans, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2014, **70**, 1248–1256.
96. S. Horrell, S. V. Antonyuk, R. R. Eady, S. S. Hasnain, M. A. Hough and R. W. Strange, *IUCrJ*, 2016, **3**, 271–281.
97. S. Bui, D. von Stetten, P. G. Jambrina, T. Prange, N. Colloc'h, D. de Sanctis, A. Royant, E. Rosta and R. A. Steiner, *Angew. Chem., Int. Ed. Engl.*, 2014, **53**, 13710–13714.
98. K. M. Polyakov, S. Gavryushov, S. Ivanova, T. V. Fedorova, O. A. Glazunova, A. N. Popov and O. V. Koroleva, *Acta Crystallogr., Sect. D*, 2017, **73**, 388–401.
99. D. Kekilli, T. Moreno-Chicano, A. K. Chaplin, S. Horrell, F. S. N. Dworkowski, J. A. R. Worrall, R. W. Strange and M. A. Hough, *IUCrJ*, 2017, **4**, 263–270.
100. R. B. Ravelli, H. K. Leiros, B. Pan, M. Caffrey and S. McSweeney, *Structure*, 2003, **11**, 217–224.
101. R. B. Ravelli, M. H. Nanao, A. Lovering, S. White and S. McSweeney, *J. Synchrotron Radiat.*, 2005, **12**, 276–284.
102. E. Rudiño-Piñera, R. B. Ravelli, G. M. Sheldrick, M. H. Nanao, V. V. Korostelev, J. M. Werner, U. Schwarz-Linek, J. R. Potts and E. F. Garman, *J. Mol. Biol.*, 2007, **368**, 833–844.
103. E. Rudiño-Piñera, U. Schwarz-Linek, J. R. Potts and E. F. Garman, *Acta Crystallogr., Sect. D*, 2004, **60**, 1341–1345.
104. M. H. Nanao and R. B. Ravelli, *Structure*, 2006, **14**, 791–800.
105. H. Yeung, C. J. Squire, Y. Yosaatmadja, S. Panjikar, G. Lopez, A. Molina, E. N. Baker, P. W. Harris and M. A. Brimble, *Angew. Chem., Int. Ed. Engl.*, 2016, **55**, 7930–7933.
106. P. Sliz, S. C. Harrison and G. Rosenbaum, *Structure*, 2003, **11**, 13–19.
107. H. K. Leiros, J. Timmins, R. B. Ravelli and S. M. McSweeney, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2006, **62**, 125–132.
108. A. Mhaisekar, M. J. Kazmierczak and R. Banerjee, *J. Synchrotron Radiat.*, 2005, **12**, 318–328.
109. U. W. Arndt, *J. Appl. Crystallogr.*, 1984, **17**, 118–119.
110. M. S. Weiss, S. Panjikar, C. Mueller-Dieckmann and P. A. Tucker, *J. Synchrotron Radiat.*, 2005, **12**, 304–309.
111. N. Shimizu, K. Hirata, K. Hasegawa, G. Ueno and M. Yamamoto, *J. Synchrotron Radiat.*, 2007, **14**, 4–10.

112. C. Homer, L. Cooper and A. Gonzalez, *J. Synchrotron Radiat.*, 2011, **18**, 338–345.
113. R. Fourme, V. Honkimäki, E. Girard, K. Medjoubi, A. C. Dhaussy and R. Kahn, *J. Appl. Crystallogr.*, 2012, **45**, 652–661.
114. D. Liebschner, G. Rosenbaum, M. Dauter and Z. Dauter, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2015, **71**, 772–778.
115. U. Zander, G. Bourenkov, A. N. Popov, D. de Sanctis, O. Svensson, A. A. McCarthy, E. Round, V. Gordeliy, C. Mueller-Dieckmann and G. A. Leonard, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2015, **71**, 2328–2343.
116. K. Hasegawa, K. Yamashita, T. Murai, N. Nuemket, K. Hirata, G. Ueno, H. Ago, T. Nakatsu, T. Kumasaka and M. Yamamoto, *J. Synchrotron Radiat.*, 2017, **24**, 29–41.
117. J. M. Holton, *J. Synchrotron Radiat.*, 2009, **16**, 133–142.
118. A. I. Barker, R. J. Southworth-Davies, K. S. Paithankar, I. Carmichael and E. F. Garman, *J. Synchrotron Radiat.*, 2009, **16**, 205–216.
119. J. Kmetko, M. Warkentin, U. Englich and R. E. Thorne, *Acta Crystallogr., Sect. D*, 2011, **67**, 881–893.
120. E. G. Allan, M. C. Kander, I. Carmichael and E. F. Garman, *J. Synchrotron Radiat.*, 2013, **20**, 23–36.
121. E. Crosas, A. Castellvi, I. Crespo, D. Fulla, F. Gil-Ortiz, G. Fuertes, C. S. Kamma-Lorger, M. Malfois, M. A. G. Aranda and J. Juanhuix, *J. Synchrotron Radiat.*, 2017, **24**, 53–62.
122. S. Boutet, L. Lomb, G. J. Williams, T. R. Barends, A. Aquila, R. B. Doak, U. Weierstall, D. P. DePonte, J. Steinbrener, R. L. Shoeman, M. Messerschmidt, A. Barty, T. A. White, S. Kassemeyer, R. A. Kirian, M. M. Seibert, P. A. Montanez, C. Kenney, R. Herbst, P. Hart, J. Pines, G. Haller, S. M. Gruner, H. T. Philipp, M. W. Tate, M. Hromalik, L. J. Koerner, N. van Bakel, J. Morse, W. Ghonsalves, D. Arnlund, M. J. Bogan, C. Caleman, R. Fromme, C. Y. Hampton, M. S. Hunter, L. C. Johansson, G. Katona, C. Kupitz, M. Liang, A. V. Martin, K. Nass, L. Redecke, F. Stellato, N. Timneanu, D. Wang, N. A. Zatsepin, D. Schafer, J. Defever, R. Neutze, P. Fromme, J. C. Spence, H. N. Chapman and I. Schlichting, *Science*, 2012, **337**, 362–364.
123. K. Hirata, K. Shinzawa-Itoh, N. Yano, S. Takemura, K. Kato, M. Hatanaka, K. Muramoto, T. Kawahara, T. Tsukihara, E. Yamashita, K. Tono, G. Ueno, T. Hikima, H. Murakami, Y. Inubushi, M. Yabashi, T. Ishikawa, M. Yamamoto, T. Ogura, H. Sugimoto, J. R. Shen, S. Yoshikawa and H. Ago, *Nat. Methods*, 2014, **11**, 734–736.
124. G. Chreifi, E. L. Baxter, T. Doukov, A. E. Cohen, S. E. McPhillips, J. Song, Y. T. Meharenna, S. M. Soltis and T. L. Poulos, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, 1226–1231.
125. L. Lomb, T. R. Barends, S. Kassemeyer, A. Aquila, S. W. Epp, B. Erk, L. Foucar, R. Hartmann, B. Rudek, D. Rolles, A. Rudenko, R. L. Shoeman, J. Andreasson, S. Bajt, M. Barthelmess, A. Barty, M. J. Bogan, C. Bostedt, J. D. Bozek, C. Caleman, R. Coffee, N. Coppola,

D. P. Deponte, R. B. Doak, T. Ekeberg, H. Fleckenstein, P. Fromme, M. Gebhardt, H. Graafsma, L. Gumprecht, C. Y. Hampton, A. Hartmann, G. Hauser, H. Hirsemann, P. Holl, J. M. Holton, M. S. Hunter, W. Kabsch, N. Kimmel, R. A. Kirian, M. Liang, F. R. Maia, A. Meinhart, S. Marchesini, A. V. Martin, K. Nass, C. Reich, J. Schulz, M. M. Seibert, R. Sierra, H. Soltau, J. C. Spence, J. Steinbrener, F. Stellato, S. Stern, N. Timneanu, X. Wang, G. Weidenspointner, U. Weierstall, T. A. White, C. Wunderer, H. N. Chapman, J. Ullrich, L. Struder and I. Schlichting, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2011, **84**, 214111.

126. A. Barty, C. Caleman, A. Aquila, N. Timneanu, L. Lomb, T. A. White, J. Andreasson, D. Arnlund, S. Bajt, T. R. M. Barends, M. Barthelmess, M. J. Bogan, C. Bostedt, J. D. Bozek, R. Coffee, N. Coppola, J. Davidsson, D. P. DePonte, R. B. Doak, T. Ekeberg, V. Elser, S. W. Epp, B. Erk, H. Fleckenstein, L. Foucar, P. Fromme, H. Graafsma, L. Gumprecht, J. Hajdu, C. Y. Hampton, R. Hartmann, A. Hartmann, G. Hauser, H. Hirsemann, P. Holl, M. S. Hunter, L. Johansson, S. Kassemeyer, N. Kimmel, R. A. Kirian, M. N. Liang, F. Maia, E. Malmerberg, S. Marchesini, A. V. Martin, K. Nass, R. Neutze, C. Reich, D. Rolles, B. Rudek, A. Rudenko, H. Scott, I. Schlichting, J. Schulz, M. M. Seibert, R. L. Shoeman, R. G. Sierra, H. Soltau, J. C. H. Spence, F. Stellato, S. Stern, L. Struder, J. Ullrich, X. Wang, G. Weidenspointner, U. Weierstall, C. B. Wunderer and H. N. Chapman, *Nat. Photonics*, 2012, **6**, 35–40.
127. K. Nass, L. Foucar, T. R. M. Barends, E. Hartmann, S. Botha, R. L. Shoeman, R. B. Doak, R. Alonso-Mori, A. Aquila, S. Bajt, A. Barty, R. Bean, K. R. Beyerlein, M. Bublitz, N. Drachmann, J. Gregersen, H. O. Jonsson, W. Kabsch, S. Kassemeyer, J. E. Koglin, M. Krumrey, D. Mattle, M. Messerschmidt, P. Nissen, L. Reinhard, O. Sitsel, D. Sokaras, G. J. Williams, S. Hau-Riege, N. Timneanu, C. Caleman, H. N. Chapman, S. Boutet and I. Schlichting, *J. Synchrotron Radiat.*, 2015, **22**, 225–238.
128. L. Galli, S.-K. Son, M. Klinge, S. Bajt, A. Barty, R. Bean, C. Betzel, K. R. Beyerlein, C. Caleman, R. B. Doak, M. Duszenko, H. Fleckenstein, C. Gati, B. Hunt, R. A. Kirian, M. Liang, M. H. Nanao, K. Nass, D. Oberthür, L. Redecke, R. Shoeman, F. Stellato, C. H. Yoon, T. A. White, O. Yefanov, J. Spence and H. N. Chapman, *Struct. Dyn.*, 2015, **2**, 041703.

CHAPTER 5

Data Quality Analysis

M. VOLLMAR*[a] AND PHIL EVANS[b]

[a] Diamond Light Source Ltd., Harwell Science & Innovation Campus, Harwell, Oxfordshire OX11 0DE, UK; [b] MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge CB2 0QH, UK
*Email: melanie.vollmar@diamond.ac.uk

5.1 Introduction

What is Data Quality? The Wikipedia definition of the ''quality'' of data is, ''Data quality refers to the level of quality of data. There are many definitions of data quality but data are generally considered high quality if 'they are fit for their intended uses in operations, decision making and planning.'''[1] The ISO 9000 (International Organization for Standardization; quality management systems standards) definition of quality is ''the degree to which a set of characteristics of data fulfills requirements. Examples of characteristics are: completeness, validity, accuracy, consistency, availability and timeliness. Requirements are defined as the need or expectation that is stated, generally implied or obligatory.''[2]

In the process of solving crystal structures by X-ray crystallography, we measure diffraction intensities from a series of images. We scale these measurements, modelling the effects of the way the experiment was done, and then merge symmetry-related and repeated observations to infer the best estimate of the intensity of each reflection I_{hkl}, along with an estimate of its standard error or $\sigma(I_{hkl})$. This is not to be confused with the standard deviation r.m.s.d., which describes the differences between the individual measurements rather than their deviation from the truth. From these

Chemical Biology No. 8
Protein Crystallography: Challenges and Practical Solutions
Edited by Konstantinos Beis and Gwyndaf Evans
© The Royal Society of Chemistry 2018
Published by the Royal Society of Chemistry, www.rsc.org

intensities, we can also calculate the structure factor amplitude $|F_{hkl}|$ and its error estimate $\sigma(|F_{hkl}|)$.

This process of scaling and merging is often called ''Data Reduction'' as it produces a shorter list of *merged* intensities (and amplitudes) from the *unmerged* list produced by the integration program. During this process, we generate a number of quality indicators which allow us to judge whether alternative processing options might be better and give us some hints about whether the data are suitable for structure determination and refinement. However, only a successful structure determination and a satisfactory refinement of the model can in the end indicate whether the data are fit for purpose.

Following the step of data reduction, a crystallographer must solve the phase problem, either by experimental phasing or by molecular replacement, in order to be able to determine the structure of a protein. Both methods require data of high quality and completeness, especially in the lower resolution range, as this information is crucial for finding the overall molecule location, shape and trace of the protein backbone. The details of side chain orientations and interactions between protein molecules or ligands, such as small molecules, peptides or DNA/RNA, are encoded by the high-resolution part of the data.

After the phase problem has been solved one has to pay attention to the details of a structure. In this structure refinement step the challenge for a crystallographer is to distinguish between true information in an electron density map, which results in atomic coordinates, and noise, introduced by experimental errors.

In X-ray crystallography, measuring data in a diffraction experiment is the last truly experimental step. Anything influencing the *accuracy* as well as the *precision* of the data influences the outcome. Any errors introduced may not be removed further downstream in the analysis and can result in findings that are hard to interpret or are simply wrong.[3,4]

5.2 Accuracy *versus* Precision and Merged *versus* Unmerged Data

In considering errors, we need to distinguish between *accuracy* and *precision*. Precision measures how well multiple measurements of the same variable (or more generally a sample) agree with each other, *i.e.* the internal consistency of the measurements. On the other hand, accuracy measures how close our measurements are to the true value, which of course we cannot know.

Figure 5.1 gives an illustration of the principles of *precision* and *accuracy*. We can estimate the precision of the measurements by fitting an error model against their scatter around the mean, usually assuming that the errors are random and uncorrelated. However, uncorrected systematic errors will lead to a reduction in accuracy, even if the precision is high (small error

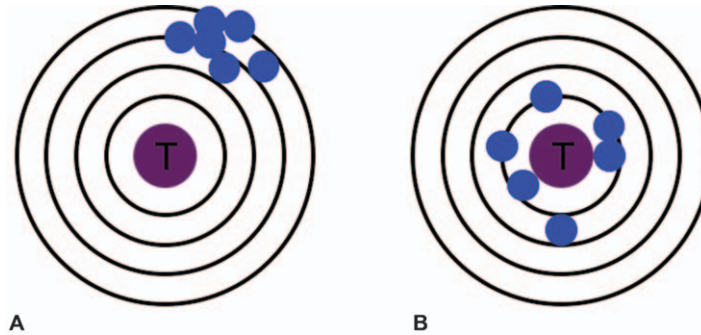


Figure 5.1 (A) Illustration of high "precision" where all repetitions of an experiment agree well with each other but low accuracy as they deviate greatly from the truth ("T"); (B) Illustration of high "accuracy" where all repetitions of an experiment are close to the truth ("T") but low precision as they deviate greatly between each other. In an optimal experiment one aims for high precision and high accuracy, *i.e.* all measurements are close to each other and are found near the truth.

estimate). Good estimation of true values requires an understanding and modelling of systematic (non-random) errors, as well as estimation of the random errors, which affect both precision and accuracy.

A series of statistical quality indicators will be discussed in the following sections to help judge the *precision* and *accuracy* of data and get an idea of their quality. We are mostly concerned with the quality of the intensities after averaging symmetry-related and repeated measurements, *merged* data, but some traditional metrics relate to the error in the *unmerged* observations, and this is sometimes a cause for confusion; statistics on *merged* and *unmerged* data cannot be compared directly, since averaging multiple observations obviously reduces the error in the mean.

5.3 Sources of Error

The difference between a measurement and what is assumed to be the truth is generally defined as error. As the truth is usually not known, its difference from the measurement cannot be determined and so any error can only be estimated based on statistical distributions, assuming the errors are random and uncorrelated. This distribution is frequently assumed to follow a Gaussian probability distribution and its estimated width describes measurement uncertainty σ.[5] In the case of diffraction experiments in protein crystallography, intensities I_{hkl} recorded multiple times for a particular Bragg reflection *hkl*, are brought to the same scale so they can be described as a Gaussian probability distribution and their associated error is given as σ_I. This σ_I will be small if the repeated measurements are distributed closely around the truth and if there is a good number of repetitions which can be used for averaging. If there are only a few or even only a single observation

for a particular reflection *hkl*, then the estimated error is usually larger, as it is difficult to estimate how close this single measurement is to the truth.

Measurements of intensities I_{hkl} differ from their true values by random errors, but also by various systematic effects arising from the experiment. Some of these systematic effects can be modelled into the corrected I_{hkl} during the scaling process and become scale-factor ''corrections'' rather than ''errors''. The estimated errors in the scale factors should be added into the total error estimate. Other effects remain as errors that we try to allow for in the estimates of σ_I. Furthermore, individual outlier measurements are an infrequent but significant source of error.

Any data integration and reduction software widely used in the protein crystallography community, such as *DIALS*,[6] *MOSFLM*,[7] *DENZO/SCALEPACK*,[8] *d*TREK*,[9] *XDS/XSCALE*[10,11] and *AIMLESS*,[12] have algorithms to determine estimates of measurement errors. They differ in their implementation and hence the results produced. It is therefore highly recommended, especially for challenging data, to try the different packages alongside to achieve the best results possible for given data.

5.3.1 Random Error

The main contribution to random error arises from the counting of photons, measured either directly with a photon counting detector, or indirectly with an integrating detector (*e.g.* CCD cameras). A photon counting detector registers the photons directly upon arrival on a pixel, whereas an integrating detector accumulates incoming photons for some time span and then reads out the summed intensity for a given pixel. The error then follows Poisson counting statistics, so for each pixel the variance of photons counted is equal to the total number of photons itself, allowing for the gain of the detector if it does not count photons directly. In the case of a photon counting detector, the gain is equal to 1. For a large number of photons N, the Poisson distribution approximates to a Gaussian distribution and an equivalent error $\sigma = \sqrt{N}$. An additional contribution to random error stems from detector read-out noise, which is negligible for photon-counting detectors but results in an increased error for CCD cameras.

The net reflection intensity is determined by subtracting a background estimate from the peak measurement, with the background under the spot derived from neighbouring pixels. For weak reflections, the error in the net intensity is dominated by the error in the background. A high background, whether arising from non-crystalline material in the beam (*e.g.* liquid around the crystal), or from detector noise, decreases the precision of the reflection intensity. If the counts in the peak area are P, and the estimated background count in the same area is B, then the net intensity $I = P - B$, and $\sigma(I) = \sqrt{(P + B)}$. Thus, in the limit for the case of $I = 0$ then $\sigma(I) = \sqrt{(2B)}$ and the peak intensity and the background cancel each other out. How reliable the determined background for a given reflection is also depends on

the way the data have been collected. With modern photon-counting detectors with negligible read-out time it is best to collect fine-sliced data (using small rotation angles), with shutterless data collection strategies, where all reflections will be recorded as partials and will be found on several adjacent images. This allows for the error in the background to be determined more precisely, resulting in a better error model and an improved signal. Figure 5.2 gives some examples of how background contributes to the diffraction signal in different levels of fine slicing. How thin the slices or how small the rotation angles can be set depends on the detector available and the mosaicity of the protein crystal. For an integrating detector, the rotation angle should not be smaller than 1/3 to 1/2 of the mosaicity, as each slice will increase the error due to the read-out noise. For photon-counting detectors, the limit is set by hardware specifications rather than the mosaicity, but there are diminishing returns in slicing reflections into more than about 5 parts (*i.e.* image width less than 1/5 total reflection width or 1/2 the XDS Gaussian width).[13]

5.3.2 Systematic Error

It is interesting to enumerate the possible sources of systematic effects or errors, so that the experiment may be designed to minimise them, or to allow their estimation in the scaling process. Variations in the experimental conditions that are much faster than the measurement time for a single reflection will be averaged automatically; for example, the pulsed time structure of a synchrotron beam is not normally a problem but might become one with extremely fast data collection. Conversely, slow variations in conditions can generally be incorporated into the scaling model. Variations in the same time domain as reflection measurement time will cause increased errors.

The model used to scale data, typically has three components: (a) a scale expressed as a function of the incident beam direction or equivalently the

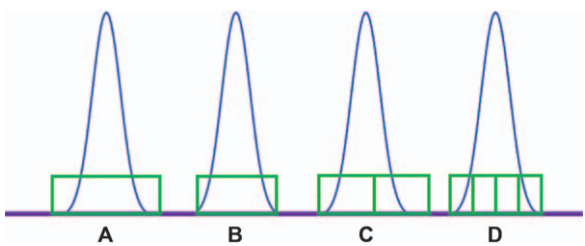


Figure 5.2 (A) Large slice/rotation angle relative to mosaicity resulting in unnecessary background being accumulated; (B) Slice/rotation angle nearly matching the diffraction spot with minimal background contribution; (C) Slice/rotation angle chosen to be slightly larger than mosaicity resulting in spots being found on more than one image but including large amounts of background as well; (D) Fine slicing, which samples a diffraction spot over several images with small background.

image number; (b) a *relative B-factor* as a function of dose, time or image number, which is largely a correction for the average radiation damage and scales up high-resolution reflections later in the exposure;[14] and (c) an absorption correction, as a function of the diffracted beam direction, often expressed as a sum of spherical harmonics.[15,16] Systematic effects can be classified according to different components of the experiment; in the list below, effects which can be modelled and therefore corrected are marked * (the estimated error in the correction factor does need to be added into $\sigma(I)$), and are related to components a, b, and c. Any systematic errors that can be corrected for by averaging over multiple experiments are marked as (d) and those that can be reduced by appropriate experiment design are indicated with (e).

The crystal and goniometer:

- variation in the illuminated volume of the crystal (* a)
- unordered packing resulting in pathologies such as twinning or split spots, secondary lattice(s)
- very long unit cell parameters (* e) or high mosaicity resulting in overlapping reflections
- radiation damage which creates a changing protein structure over time (* b, see below)
- non-isomorphism between different parts of a crystal or between different crystals of the same protein (even when harvested from the same crystallisation drop), when multiple crystals are merged
- absorption by the crystal itself and by the crystal mount (loop), air and other equipment which may be introduced while collecting rotation images (* a and c)
- vibrations of the crystal in the beam, changing the exposed volume and the orientation – often caused by a floppy loop moving in the cryostream (a)
- bad centring of the crystal, *e.g.* the influence of gravity on a crystal with a horizontal rotation axis (mitigated by a vertical axis) (a)
- anisotropy of the crystal where diffraction along one or two unit cell axis is significantly better or worse than in the other(s)
- lattice modulation
- Lorentz factor; for reflections close to the Ewald sphere, this varies rapidly, and possibly varies across different parts of the same reflection
- variation in rotation rate due to goniometer imperfections (within one or a few images)

The beam and optics:

- fluctuations in beam, at a time scale comparable to reflection measurement time
- uniformity of beam
- divergence in the beam

- shutter jitter and synchronisation, not a problem for shutterless collection with modern fast detectors
- imperfections in optics equipment (* d)

The detector:

- detector calibration for overloads, pixel position, hot pixels and sensitivity (each detector is unique and needs a description) (* e)
- inconsistent pixel size
- saturation of photon counting (* e)
- tile and panel boundaries (* e)
- shadows from other hardware (* e)
- distance to sample which can increase the background (* e)

The background estimation:

- ice or salt diffraction rings – background estimation is difficult with uneven background (some *, d)
- ice or salt spots – identified as outliers (*)
- zingers – identified as outliers (*)

The software:

- wrong assumptions about the experiment geometry and as a result creating the wrong scaling and error models

The things you may not have thought of:

- a busy road with lots of lorries and buses going past the lab where the X-ray machine is located resulting in vibrations
- machines operating in the synchrotron building you are unaware of

Most synchrotrons and hardware manufacturers attempt to reduce any hardware errors as much as possible and where this is not possible the various software programs listed above try their best to model and describe any error unaccounted for. Also, the way in which diffraction data have been measured and the experiment has been planned and conducted have an influence on how much of this modelling actually needs to be done. In Section 5.4.5 we will discuss in more detail how one can try to reduce systematic error.

5.3.3 Outliers

Outliers can be identified as they deviate from other usually symmetry-related observations much more than would be expected based on their estimated experimental error.[17] The most common cause of outliers is ice spots.

Zingers, which are random events created by cosmic rays, are another example. All data scaling programs identify and reject such outliers, so they do not get used in down-stream analysis; clearly this is easier to do with high multiplicity data where symmetry-related observations are less likely to be influenced by the occurrence of identical problems. An excessive list of outliers identified in the data integration step, *e.g.* more than five rejected reflections per integrated image, should be considered suspicious and requires investigation. Systematic error that has not been accounted for can sometimes be misinterpreted as outliers. The program *AIMLESS*[12] plots the position of outliers on the detector, which visualises problems caused by ice rings or beam-stop shadows.

5.3.4 Radiation Damage

Radiation damage is the dominating systematic error in data collected at third generation synchrotrons, which have beams that can destroy a crystal in seconds even at cryo-temperatures of 100 K (see Chapter 4). Currently there is no method available which can model and describe this type of error in full, as the underlying molecular structure is changing during the experiment. Proper treatment would need a dose-dependent refinement of changing models, using unmerged data, something that current refinement programs cannot do yet.

In the course of a diffraction experiment, an increasing dose of the X-ray beam is deposited in the protein crystal. As ionising radiation, X-rays get absorbed by the crystal and gradually some of the chemical bonds in a protein molecule will break. Radiation damage also causes a decrease in the crystalline order, hence a decay in the intensities of high-resolution reflections. There are two major reactions involved in this process. The first is the primary reaction whereby the X-ray beam directly but unspecifically interacts with electrons in the protein. It affects all atoms in a protein crystal and usually manifests in altered occupancies especially for the heavier atoms such as sulfurs or intrinsic metals.[18] Additionally, there are secondary reactions, which occur when the X-ray beam creates free radicals and electrons. These reactive species then diffuse within the crystal, even at cryo-temperatures of 100 K, and attack the protein molecules. The different amino acids in a protein react differently to the effect of radiation damage, with cystines (disulfide bridges), aspartic acid, and glutamic acid being especially susceptible to secondary reactions. This localised radiation damage often occurs before the global diffraction limit of the crystal has been reached and requires careful inspection of the electron density map. In some cases, it has been shown that the addition of radical scavengers, such as ascorbate, can help to reduce secondary radiation damage.[18,19] Radiation damage is an inevitable side reaction of the experiment and its effects can only be reduced but not avoided.[20–26] More recently, the use of free-electron lasers attempted to outrun radiation damage by collecting single images of

submicron crystals on a time scale much faster than radiation damage can occur.

When merging the intensities, we create a time average not taking into account that in fact the structure present in the crystal at the beginning of the experiment is different from that at the end. Ravelli and McSweeney (2000)[20] give some intriguing examples where they show specific structural changes in models derived from successive data sets collected from the same crystal.

Heavy atoms, whether intrinsic or introduced for experimental phasing, have greater absorption for X-rays than the surrounding protein and mother liquor, so they enhance radiation damage, and may show reducing occupancy during exposure. This is particularly a problem with the X-ray wavelength tuned on the absorption edge to maximise anomalous scattering. For heavy atoms introduced by soaking it may be useful to back-soak the protein crystal before cryo cooling to remove any excess scatterers from the surrounding liquid, which will not contribute to the anomalous signal but will increase radiation damage.

5.4 Estimating Errors

The scaling process brings the data to the same scale and makes them internally consistent, maximising the agreement between symmetry-related observations.[27,28] After all the data are on the same scale, one can start evaluating their quality and determine the errors. As with data integration, a variety of software packages are available for scaling, such as *AIMLESS*,[12] *SCALEPACK*[8] and *XDS/XSCALE*.[10,11]

5.4.1 Estimation of $\sigma(I_{hkl})$

Integration programs provide an initial estimate of $\sigma(I_{hkl})$, based primarily on counting statistics (see above). However, especially for integrating detectors, this is nearly always an underestimate of the true error, as many uncorrected fluctuations have a proportionately larger effect on strong reflections. This proportionality is most obvious from thinking about random fluctuations in the incident beam intensity, but random vibrations and deviations from constant rotation rate have a similar effect. Scaling programs therefore inflate the $\sigma(I_{hkl})$ values, typically using an error model which depends only on intensity, $\sigma'^2(I) = a(\sigma^2(I) + bI^2)$ in *XDS*[10] or $\sigma'^2(I) = SdFac^2(\sigma^2(I) + SdB{\times}I + (SdAdd{\times}I)^2)$ in *AIMLESS*.[28] In these models, the multiplying parameters a or *SdFac* mainly arise from a possible non-unit gain of the detector; if a (true) photon count of c produces a detector reading of ν, then the gain $g = \nu/c$, with the Poisson error in ν being $\sigma(\nu) = \sqrt{g}\sqrt{\nu}$, *i.e.* ideally $a^2 = SdFac^2 = g$. Parameters b and *SdAdd* model the intensity-dependent fluctuations. The parameter *SdB* has no obvious physical meaning and ideally should be close to zero. The parameters (a, b) or (*SdFac, SdB, SdAdd*) are adjusted to make the average variance $\sigma'^2(I)$ equal to the average

squared scatter of observations. The correctness of the $\sigma'(I)$ values can be judged using the reduced chi-squared statistic χ^2 (https://en.wikipedia.org/wiki/Reduced_chi-squared_statistic), analysed against intensity (and potentially also resolution and image number); this statistic should be ~1.0 if the error estimates are correct.

$$\chi^2_\nu = \frac{1}{N_{\text{free}}} \sum_h \sum_l \frac{(I_{hl} - \langle I_h \rangle)^2}{Var(I_{hl})} \tag{5.1}$$

for reflections measured at least twice, where N_{free} = (Number of observations) − (Number of unique reflections). The usual error model described above does not account for all the possible errors, in particular it does not allow for radiation damage, often the most serious error. However, analysis of the residual errors can indicate how the data collection protocol could be improved, *e.g.* by reducing X-ray exposure to minimise radiation damage.

5.4.2 *ISa*, an Indicator for Systematic Error

A detailed explanation of *ISa* (asymptotic signal-to-noise ratio) can be found in Diederichs (2010).[29] The error model described in Section 5.4.1 implies a maximum signal-to-noise ratio, $I/\sigma(I)$ (before merging), for large intensities, when the intensity-dependent term dominates.

If $\sigma'^2(I) = a\ (\sigma^2(I) + b\ I^2)$, the asymptotic value of $I/\sigma'(I)$

$$ISa = (I/\sigma)_{\text{asymptotic}} = \frac{1}{\sqrt{(ab)}} \approx \frac{1}{SdFac \cdot SdAdd} \tag{5.2}$$

For synchrotron data, the upper limit for $I/\sigma(I)$ was found to be *ISa* ~30 for CCD cameras and *ISa* ~60 for photon-counting detectors. Individual reflections, especially at low resolution, can exhibit 1000 to 10 000 counts, so a good crystal should allow for *ISa* > 100, tending to infinity if only random error is present. Diederichs (2010)[29] suggested that unmodelled errors from the experiment were reducing the accuracy of the data. *ISa* is derived from plotting the intensity $I/\sigma(I)$ against I (number of counts), or from eqn (5.2). Provided that the $\sigma(I)$ estimates are good, a large value of *ISa* indicates a good data set, and it was suggested that *ISa* >25 will allow structure solution by experimental phasing. This was supported by plotting *CCanom*[30] *versus ISa*, which showed that both are affected by systematic error in the same way. If the error model used during integration and scaling/merging is ill-defined, then *ISa* is no longer informative.

5.4.3 R_{merge}, R_{sym} and R_{meas} – Indicators for Unmerged Data

In 1968 Arndt *et al.*[31] introduced R_{merge} (otherwise known as R_{sym} or R_{int}) in an attempt to expand the concept of *R factors* (R_{cryst}) introduced by Wilson

(1950)[32] for refinement, to the process of data reduction. R_{merge} is useful when judging the average precision of individual observations of intensities. However, it does not account for multiplicity (observing the same reflection repeatedly) and as a result it makes data with low multiplicity appear to be of better quality than data with multiple observations for each reflection, although clearly increasing multiplicity improves the precision of the average. Thus, R_{merge} is not a good measure to judge the overall quality of a data set. However, it is quite a useful indicator to identify outlier images in a data set when monitored over rotation range/image number/time.[27]

$$R_{\text{merge}} = R_{\text{sym}} = \frac{\sum_h \sum_l |I_{hl} - \langle I_h \rangle|}{\sum_h \sum_l \langle I_h \rangle} \; or \; \frac{\sum_h \sum_l |I_{hl} - \langle I_h \rangle|}{\sum_h \sum_l I_{hl}} \tag{5.3}$$

These alternative denominator definitions (and their equivalents for related *R-factors*) are used in different programs, but in most cases, they will give similar values.

To take multiplicity into account another indicator has been suggested, R_{meas}[33] or $R_{\text{r.i.m.}}$.[34] Both *R-factors* look at all the observations for a particular reflection (weighted average value $\langle I_{\text{h}} \rangle$), and measure how well they agree with each other. The crystallographic community has been slow to adopt the improved R_{meas} to replace R_{merge}, probably because its resulting values are larger.

$$R_{\text{meas}} = R_{\text{r.i.m.}} = \frac{\sum_h \sqrt{\frac{n_h}{n_{h-1}}} \sum_l |I_{hl} - \langle I_h \rangle|}{\sum_h \sum_l I_{hl}} \tag{5.4}$$

Krojer and von Delft (2011)[35] pointed out the relation between R_{meas} and $\langle I/\sigma_I \rangle_{\text{obs}}$ is $R_{\text{meas}} \sim 0.8/\langle I/\sigma_I \rangle_{\text{obs}}$, as long as the error model used is well defined, and the internal consistency is given by $\chi^2 \sim 1$. Note that even with the multiplicity correction, R_{meas} refers to the error in the *unmerged* data.

5.4.4 $R_{\text{mrgd-}I}$, $R_{\text{p.i.m.}}$, R_{anom} and $CC_{1/2}$ – Indicators for Merged Data

For virtually all downstream analysis in crystallography, *i.e.* structure solution and refinement, merged data are used, so it is useful to consider metrics which judge the quality of data after merging.[36] Two indicators which take multiplicity into account have been suggested to describe the quality of merged data, $R_{\text{mrgd-}I}$[33] and $R_{\text{p.i.m.}}$.[34] $R_{\text{mrgd-}I}$ is the same as R_{split} commonly used in the free-electron laser community and looks at the differences of

merged intensities between randomly chosen subsets of data. $R_{\text{p.i.m.}}$ also known as ''precision indicating merging'' *R-factor* is derived from R_{meas}.

$$R_{\text{p.i.m.}} = \frac{\sum_h \sqrt{\frac{1}{n_{h-1}}} \sum_l |I_{hl} - \langle I_h \rangle|}{\sum_h \sum_l I_{hl}} \tag{5.5}$$

$$R_{\text{mrgd-}I} = \frac{\sum \left| I_{h,P} - I_{h,Q} \right|}{0.5 \times \sum I_{h,P} + I_{h,Q}} \tag{5.6}$$

with $I_{h,P} = \dfrac{1}{n_{h,P}} \sum_{i \in P}^{n_{h,P}} I_{h,i}$ and with $I_{h,Q} = \dfrac{1}{n_{h,Q}} \sum_{i \in Q}^{n_{h,Q}} I_{h,i}$.

Quite frequently comparisons are made between $R_{\text{p.i.m.}}$ and R_{anom}. The latter describes the precision of the anomalous signal and was introduced by Weiss (2001).[34]

$$R_{\text{anom}} = \frac{\sum_{hkl} \left| I(hkl) - I\left(\bar{h}\bar{k}\bar{l}\right) \right|}{\sum_{hkl} \overline{I(hkl)}} \tag{5.7}$$

which compares Friedel pairs rather than symmetry-related observations. As with $R_{\text{meas}} \sim 0.8 \langle I/\sigma_I \rangle_{\text{obs}}$ for unmerged data there is a reciprocal relationship between $R_{\text{mrgd-}I}$ and $R_{\text{p.i.m.}}$ and $\langle I/\sigma_I \rangle_{(\text{mrgd})}$. The exact ratio is dependent on the integration program used and how the errors have been estimated.

The most recent, and most useful, addition to the family of precision indicators for merged data is $CC_{1/2}$, calculated in resolution shells. This also takes multiplicity of the data into account, and is defined as the Pearson's correlation coefficient between two randomly chosen half-sets of merged intensities.[12,36,37]

$$CC_{1/2} = \frac{\sum (a_i - \bar{a})(b_i - \bar{b})}{\left[\sum (a_i - \bar{a})^2 \sum (b_i - \bar{b})^2 \right]^{\frac{1}{2}}} \tag{5.8}$$

where a and b are intensities from two randomly chosen half-datasets. A similar metric, the Fourier Shell Correlation, with complex rather than real coefficients, is widely used in the cryo-electron microscopy community. By definition $CC_{1/2}$ is independent of standard deviations and has well-understood statistical properties, unlike *R-factors*, so that tests such as *Student's t-test*[38] may be used to check the data for its significance, and hence can be used to determine a high-resolution cut-off at the point where there is no longer any useful signal available. The correlation coefficient is a normalised covariance, *i.e.* $CC(a,b) = Covariance(a,b)/\sqrt{[Var(a)Var(b)]} = Covariance(a,b)/Var(a)$, since the half-datasets are equivalent, thus $Var(a) = Var(b)$. Assmann, Brehm and Diederichs (2016)[39] point out that $CC_{1/2}$ can

be calculated (in resolution shells) without explicit splitting of the data, as

$$CC_{1/2} = \frac{\sigma_y^2 - \sigma_{\bar{\varepsilon}f}^2}{\sigma_y^2 + \sigma_{\bar{\varepsilon}f}^2} \tag{5.9}$$

where σ_y^2 is the variance of $\langle I(hkl)_{\mathrm{obs}}\rangle$, *i.e.* the dispersion of observed intensities, and $\sigma_{\bar{\varepsilon}f}^2$ is the average sample variance of the individual observations ($\sigma_{\bar{\varepsilon}f}^2 = \frac{1}{2}\sigma_{\varepsilon}^2$, the average variance for each half dataset). This derivation assumes that the errors are random and uncorrelated, and also explains how negative $CC_{1/2}$ (and indeed negative CC_{anom}) can arise, if the dispersion of individual observations is greater than the dispersion of true intensities.

5.4.5 R_{d}, R_{cum} and *B Factor* – Indicators for Radiation Damage

A number of metrics have been suggested for monitoring radiation damage as a function of image number or crystal rotation (related to X-ray dose). During data collection, many beamlines monitor the number of spots found, or total intensity, particularly for high-resolution reflections, against the absorbed dose.[40] In the data reduction process, a number of statistics are typically plotted against image number (or groups of images); these include R_{merge}, which will increase with radiation damage; a maximum resolution estimate; and the relative *B-factor* (''temperature factor''). The latter is often part of the scaling model as $\exp(-2B_i \sin^2\theta/\lambda^2)$ with B_i as a (smooth) function of image number – a negative *B-factor* scales up high-resolution reflections, so a large negative value indicates that the scaling function is compensating for a radiation-induced suppression of high-resolution terms.

Alternative metrics include the 'decay R factor', R_{d}, as suggested in Diederichs (2006)[41]

$$R_{\mathrm{d}} = \frac{\sum_{hkl}\sum_{|i-j|=d}\left|I_i - I_j\right|}{\sum_{hkl}\sum_{|i-j|=d}\frac{\left(I_i + I_j\right)}{2}} \tag{5.10}$$

for intensities from images i and j, or a related 'cumulative R-factor' suggested by Graeme Winter (2009)[42]

$$R_{\mathrm{cum}}(\mathrm{Max}(i,j)) = \frac{\sum_{hkl}\sum_{i,j}\left|I_i - I_j\right|}{\sum_{hkl}\sum_{i,j}\frac{\left(I_i + I_j\right)}{2}} \tag{5.11}$$

Diederichs[41] (2006) also suggested correcting for radiation damage for each reflection, by interpolating or extrapolating each series of equivalent

observations to a zero-dose or common dose-time-point.[43] However, this process is difficult, as it depends on high multiplicity and a good time distribution of observations, and also assumes that the X-ray beam matches the crystal size and hence there will always be the same volume of crystal irradiated. If the crystal is bigger than the beam then fresh material is introduced into the beam during rotation, which makes modelling the decay more complicated.

5.4.6 Data Completeness

Incomplete data leads to distortions in the electron density maps, which makes them harder to interpret. Some graphical examples to illustrate the effect of missing data on the electron density map can be found on the following urls (http://bl831.als.lbl.gov/~jamesh/movies/completeness.mpeg; http://www.ysbl.york.ac.uk/~cowtan/fourier/duck4.html). There are various reasons why completeness may be systematically compromised including exclusion of ice rings or other shadows on a diffraction image, a detector position offset from the beam centre, gaps between detector tiles, strong low-angle reflections overloading the detector and radiation damage forcing the rejection of the later parts of a data collection. The level of incompleteness that results in a still interpretable electron density has never been tested in detail. Random missing reflections are unlikely to cause problems, but exclusion of a large contiguous volume of reciprocal space causes convolution of the electron density maps with the transform of the missing volume, and thus seriously distorted density. Features such as non-crystallographic symmetry (NCS) can help to compensate a lack of data completeness as it allows averaging over multiple copies of the same molecule in the asymmetric unit, but the aim should always be to collect complete data.

5.5 Use of Metrics

This section aims to provide some guidance on how to get the most out of your data and how one may use some of the metrics introduced above to judge the quality of given data. Applying those guidelines, a crystallographer should be able to make an informed decision on how internally consistent the data is, if the error model used is appropriate, if any data need to be excluded for valid reasons other than ''data massaging'', how complete the data are and what the ''resolution'' can be considered to be.

5.5.1 BLEND and Merging Multiple Crystals and/or Data Sets

With small weakly diffracting and radiation-sensitive crystals, it may not be possible to collect a complete dataset from a single crystal before it is destroyed by radiation damage. However, merging data from multiple crystals requires testing for isomorphism, *i.e.* that they are essentially identical.

BLEND[44] is a program that can help a crystallographer to combine multiple data sets. It was developed in the course of solving the structure of a membrane protein, whose crystals were rather small in size and additionally very sensitive to radiation. After collecting 20 data sets of 25° to 60° rotation at a microfocus beamline with a high-intensity beam at a third generation synchrotron, none of them individually provided enough information to solve the phase problem and determine the protein structure.

The principal of combining multiple datasets has been applied in the past for room-temperature data collection, but it can be a very time-consuming process to assess all possible combinations to obtain a complete, good quality dataset, that allows for structure solution. *In situ* room-temperature data collection, with crystals staying untouched in their crystallisation plate, is becoming increasingly popular and at some synchrotrons, *e.g.* Diamond Light Source, UK, beamlines are built to be entirely dedicated to measure this type of data. In-plate data collection will only allow for a small wedge of data, possibly only a few images, to be collected of each crystal within the plate and will make using programs like *BLEND*[44] a default step in the analysis.

BLEND[44] semi-automatically tests a large number of data sets (even from single images recorded as stills) for non-isomorphism, ranks and clusters them and then uses the programs *POINTLESS* and *AIMLESS* in an attempt to create the best possible dataset. The data used in the analysis stems from the data integration step and hence has not been scaled and merged. The various datasets are analysed with multivariate statistics, in particular cluster analysis, to group them for similarity, based on the differences in unit cell parameters. The smaller the differences the more likely those datasets contain the same structure and can be combined. The quality of the resulting *merged* data set is then judged in the same way as for a single sweep data collection.

Overall, combining multiple data sets can reduce the systematic errors found in the data (by way of averaging observations with different errors), and increases the multiplicity as well as the completeness.

5.5.2 Identifying Rogue Data Sets and Linking Data and Model Quality

Recently, Assmann, Brehm and Diederichs (2016)[39] published another approach to selecting which datasets to combine, especially data comprised of very small image ranges, which is based on $CC_{1/2}$. In this method, for each dataset $CC_{1/2}$ is first calculated in resolution shells, then averaged over the shells and then over all datasets to get $CC_{1/2_\mathrm{overall}}$. Next, they derive $CC_{1/2_\mathrm{overall}}$ excluding the *ith* dataset as $CC_{1/2_i}$, and then determine the incremental correlation for each individual data set as:

$$\Delta CC_{1/2_i} = CC_{1/2_i} - CC_{1/2_\mathrm{overall}} \tag{5.12}$$

If $\Delta CC_{1/2_i} > 0$ then the *ith* dataset improves the overall data, whereas if $\Delta CC_{1/2_i} < 0$ it has a deteriorating effect on the overall data and should be excluded. This process can then be iterated for each wedge of data. In their examples, a histogram of the different $\Delta CC_{1/2_i}$ showed most data clustered around 0, *i.e.* inclusion of that dataset made little difference, but a few outliers which were either deleterious or particularly useful can be identified as well. Furthermore, the authors examined the correlation between $\Delta CC_{1/2_i}$ and an analogous correlation with intensities calculated from the refined model $\Delta CC_{\mathrm{FOC}_i}$, showing that $CC_{1/2}$ is a good metric for judging the quality of data. A further extension of this idea is to separate systematic and random errors using a multivariate analysis of the inter-dataset correlation coefficients, and to use this analysis to classify multiple datasets into isomorphous groups.[45] $CC_{1/2}$ is therefore a reliable indicator when estimating the correlation between the observed merged intensities $\langle I_{hkl} \rangle$ and their (unknown) true values, $CC_{\mathrm{true}} \approx CC^*$,[37]

$$CC_{\mathrm{true}} \approx CC^* = \sqrt{\frac{2CC_{1/2}}{1 + CC_{1/2}}} \tag{5.13}$$

This CC_{true} or CC^* can then be compared to the correlation coefficient between the observed data and structure factors F^2_{calc} calculated from an atomic model.[36,46] The resulting correlation coefficient can then be split into CC_{work} and CC_{free}, similarly to R_{work} and R_{free} used during refinement of atomic coordinates. CC_{work} and CC_{free} should converge to CC^* during the final rounds of refinement. Values greater than CC^* suggest overfitting and may result in building atoms into noise of the electron density map. Any improvement in data quality, for example due to improved algorithms in data processing or by including more data from higher resolution shells or additional datasets, should therefore produce an increased CC^* which in turn would create a higher CC_{work} and CC_{free} with decreased R_{work} and R_{free} in refinement.[36] R_{merge} is not suitable to be used in comparison with R_{cryst} values, since with weak data R_{merge} tends to infinity, while R_{cryst} tends to a constant (about 0.43).[12]

5.5.3 Determination of a High-resolution Cut-off

Observed intensities become weaker towards higher resolution and, at some point, adding another resolution shell of data will add no significant information and may even degrade the dataset. Judging where to cut the data has been a contentious issue, particularly when refereeing articles which quote a particular resolution for a structure. Historically, a number of criteria have been suggested (though rarely documented), and the choice of a cut-off point has not been helped by the assumption that a single number, the ''resolution'' of the data, can be taken as an indicator for the quality of the model. The worst metric that has been used for setting a resolution

cut-off is R_{merge}, both because it measures the error in the *unmerged* data, while it is the error in the *merged* intensities (or amplitudes) which affect the model, and also because it tends to infinity for weak high-resolution data (as does *e.g.* $R_{\text{p.i.m.}}$), so there is no sensible cut-off point. A better measure is $\langle I/\sigma_I \rangle_{\text{obs}}$, with a cut-off variously recommended as between 1 and 2. But estimates of σ_I are not always accurate (see above). However, it is now clear that the best metric is $CC_{1/2}$, perhaps cutting the data at values somewhere between 0.1 and 0.5. Low-resolution data usually has $CC_{1/2}$ ~1.0 and with increasing resolution this drops off to 0.0 ($CC_{1/2}$ much less than 1.0 at low resolution indicates a serious problem in the data).

One complicating factor for determining a high-resolution cut-off is the presence of anisotropy. Most crystals show anisotropic diffraction to some extent, except in cubic space groups, so the collected data will have varying resolution limits along their principal unit cell axes. This often changes in the course of data collection and can be seen to varying degrees in the diffraction images (see Figure 5.3 for examples). The differences can range from fractions of, to several Ångström units. Anisotropic resolution cut-offs, for example if data has been submitted for anisotropy correction to the ''Diffraction Anisotropy Server'' (https://services.mbi.ucla.edu/anisoscale/)[47] or ''The STARANISO Server'' (http://staraniso.globalphasing.org/server_about.html; Global Phasing Limited) often result in artefacts being introduced in the final electron density map as the data are incomplete, and should be avoided. With current refinement methods, it is unclear what is the best way of handling anisotropic data.

The effective resolution of an electron density map varies for different parts of the structure, depending on the degree of disorder (measured roughly as the ''temperature factor''; this disorder can be a result of radiation damage or it is based on intrinsic flexibility of the protein within the crystal), and may be greatly improved if non-crystallographic symmetry, NCS, is present and one can average over multiple molecules in the asymmetric unit.

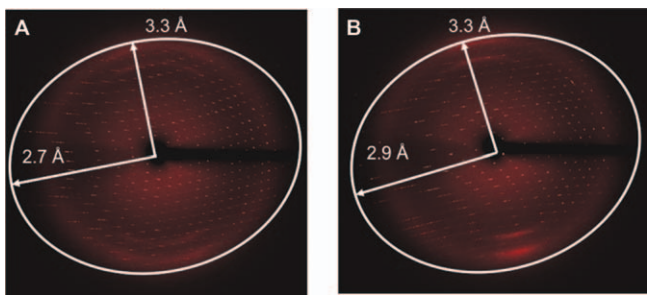


Figure 5.3 Example anisotropy in diffraction images; (A) and (B) are images from within the same data set. (A) was taken at a rotation angle of 271° and (B) at 321°, 50° apart. Both images show anisotropic diffraction with the resolution along two of the primary axes differing by 0.6 Å in (A) and 0.4 Å in (B) (images taken with dials.image_viewer6).

A practical approach to decide the high-resolution limit is using a ''paired-refinement'' strategy at the end of model building. In this method, the same model is refined successively against data with two increasing resolution limits, and R_{free} calculated for the two refined models against the lower resolution set, *i.e.* a better model will improve the original R_{free} and is not just fitting noise in high-resolution shells. The ''best'' resolution limit is then the point at which adding the next shell of data does not improve R_{free} or will fail to resolve more details in a model.[12,36,37]

5.5.4 Things to Consider When Collecting and Analysing Data

The overall goal of any diffraction experiment is to get the best diffraction with the lowest possible error, random and systematic as well as outliers, for the type of experiment conducted. This usually requires some trade-offs and may involve opposing procedures, *e.g.* collect near the absorption edge of a particular anomalous scatterer to increase the anomalous signal for experimental phasing at the risk of severe radiation damage and possibly high incompleteness. Before starting any data processing (or even before collecting a full set of images) it is advisable to examine several randomly picked diffraction images visually, which may immediately indicate that some crystals are better than others. Commonly two or three test images 90° or 45° apart are collected to assess the diffraction potential of a crystal. This already provides an indication whether there may be problems with anisotropy (different resolution limits for the different unit cell directions) or whether the spots have pathologies, *e.g.* blurred smeary spots, or multiple lattices. To get an idea of how some of these pathologies manifest in diffraction images, some examples are given in Figure 5.4. Large unresolved spots, or high mosaicity are likely to lead to poor, incomplete data, due to spot overlap.

Software packages usually offer some tools to model the pathologies. Multiple lattices are a fairly common problem but are not very well handled yet with current software. *MOSFLM*[7] can index and integrate multiple lattices in favourable cases (for well resolved lattices), and multiple lattices can be indexed in *DIALS*,[6] though their proper integration is not yet available. However, currently there are no refinement programs which handle data from partly overlapped lattices (*e.g.* non-merohedral twin) where the lattice ratio varies with crystal rotation (*SHELXL* can refine such data, with the assumption that the lattice ratio is constant).

Ice rings are another common problem; for data collected with a fine-slicing method on photon-counting detectors ice rings (see Figure 5.5) may not be obvious visually, but a new tool *AUSPEX* can detect them in the integrated intensities,[48] so that contaminated data may be excluded (at the cost of losing part of reciprocal space).

If symmetry-related reflections are measured in a way such that uncorrected systematic errors are different, then averaging over multiple

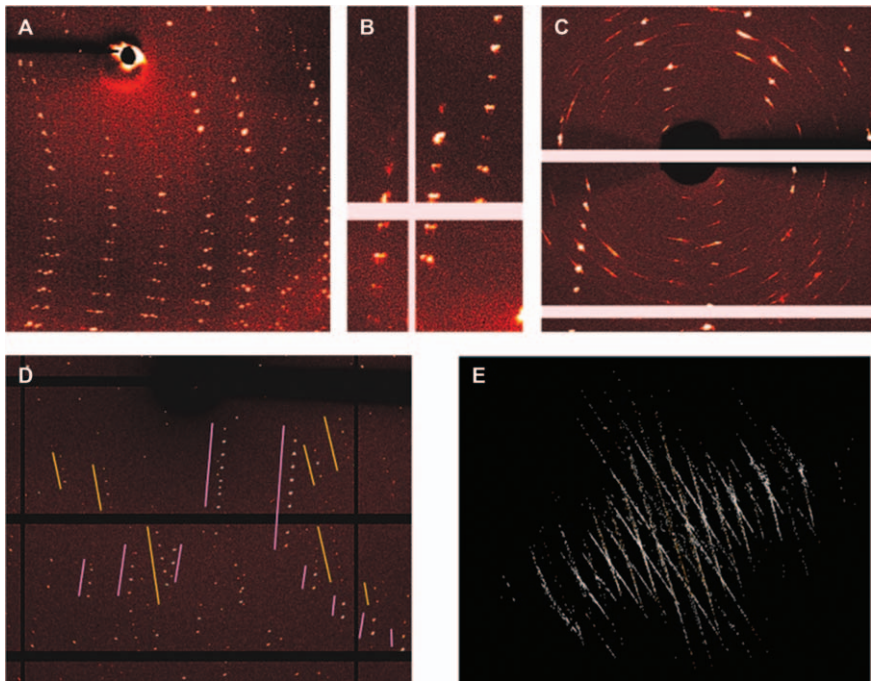


Figure 5.4 Examples of lattice pathologies: (A) satellite spots indicating some lattice modulation; (B) split spots with increasing separation the higher the resolution indicating twinning; (C) smeared out spots indicating a circle which looks similar to powder diffraction and points to multiple crystals; (D) and (E) represent multiple lattices, in (D) the two major lattices are indicated by pink and orange lines, (E) representation of two lattices in dials.reciprocal_lattice_viewer[6]; images (A) to (D) have been generated with dials.image_viewer[6].

observations will result in each measurement being closer to the truth, and the averaging will reduce the effect of random errors as well. This can be achieved for example by collecting data over multiple revolutions with the crystal orientation changed for each rotation, by using a mini-*kappa* goniometer[49] or a SmarGon (SmarAct GmbH), which can be accessed at many synchrotrons, or alternatively by manually bending the crystal mount. If done using low transmission and therefore low dose, radiation damage can be minimised. With modern fast zero-noise detectors, the best strategy is probably to collect a large amount of data quickly.

Detectors generally need calibration, and this is done by the manufacturers or by the staff at synchrotron beamlines, but some aspects of spatial distortion may be modelled and refined in the integration process. Other systematic errors are only corrected in the scaling process if they are modelled and if errors are different for the symmetry-related observations, since the scaling process can only make the data internally consistent, not

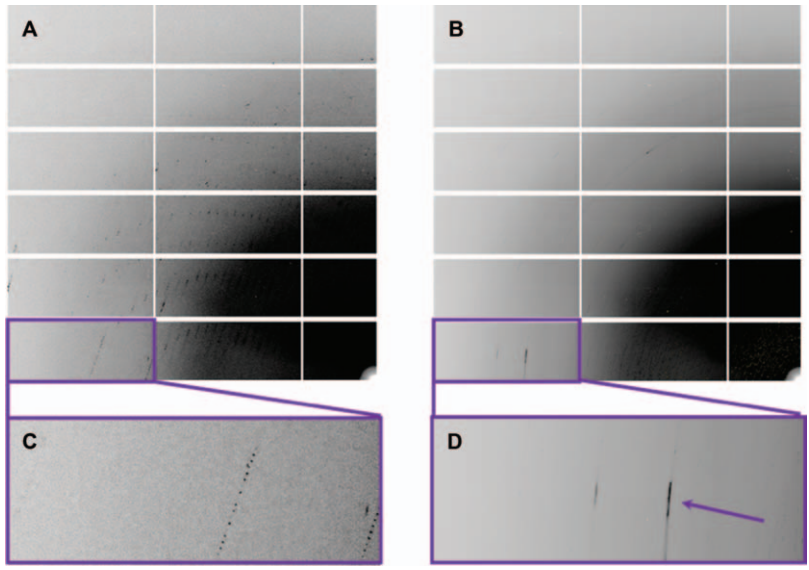


Figure 5.5 Ice rings in fine-sliced data on photon counting detector; (A) single image of a fine-sliced data collection on a photon counting detector with no obvious ice rings or ice spots; (B) the same data set but creating a sum over all images, here 675, with now appearing ice rings (indicated by a purple arrow in zoom-window (D) with no ice-rings to be seen in zoom-window (C)) (images created with dials.image_viewer and summation carried out with dials.merge_cbf[6]).

necessarily accurate. Thus, for the most accurate data, the collection strategy should be to collect equivalent reflections in different ways, so that the average might be closer to the truth, even if this makes the consistency statistics worse. The extreme example of this is the massive multiplicity averaging used in serial crystallography (''Monte Carlo integration'').[50]

Integration and scaling applies a series of models, which should match and describe the experiment. An example of such modelling is the estimation of the background under a peak, which is subtracted from the recorded peak to get the net intensity. An inaccurate model may however result in a wrong estimation of the background, such as where an ice ring causes a highly variable background.[51,52]

It is also worth noticing that strong and weak reflections are influenced differently by systematic and random error. The accuracy of weak reflections is dominated by errors in the background, mainly random errors. On the other hand, strong reflections are mostly affected by instrument fluctuations. Diederichs (2010)[29] suggests using *ISa* as a measure to identify systematic error when comparing the results of different datasets collected from the same or different crystals.

Detector overloads are a less important problem with modern detectors than in the past, but it may still be useful to collect an additional set of

rotation images with reduced transmission to avoid a large number of overloads, which will be rejected in the data reduction step, and to avoid reduced low-resolution completeness. At synchrotrons, one can often find pre-set positions for beamstops that allow the collection of the reflections at the very lowest resolution.

Radiation damage is probably the most important systematic error, because averaging over multiple observations will not reduce the error. It may be necessary to combine data from different crystals, but this requires the crystals to be isomorphous. A number of statistical methods have been developed to find the data combination that yields the best data set and several examples for their application can be found in the literature.[44,45,53–58]

When collecting data for experimental phasing (SAD, MAD or S/MIRAS), it is important that the reflections which will be used to determine anomalous differences (often called $I+$ and $I-$, or $F+$ and $F-$) are measured in as similar a way as possible, in particular close together in time, so that radiation damage does not affect the measured difference. This can be done either by rotating approximately around a crystallographic 2-fold axis (maybe using a mini-*kappa* goniometer), or with an inverse beam strategy. Most synchrotron beamlines have such strategies available.

5.6 Concluding Remarks

In this chapter, we presented the background to understanding, and practical guidelines to improve, data quality. As the collection of diffraction data is the last experimental step, utmost care should be given to its execution and the list below may be used as a reference to aid experiment design:

1. Harvest crystals with the loop matching the size of the crystal; backsoak heavy atom solutions; add cryo protectant and/or radiation scavenger; this helps to reduce the background
2. Assess the crystal with some low-dose test images; determine diffraction strength, crystal symmetry, cell dimensions
3. Optimise beamline geometry for your crystal; detector position to reduce background; alignment along symmetry axes to avoid overlaps and collect anomalous pairs; match the beam size to the crystal; ask beamline staff for help
4. Collect multiple revolutions of fine-sliced data at low-dose with varying starting orientations; amend for anomalous data collection and integrating detectors accordingly
5. Combine data sets from multiple sweeps and/or crystals; watch for radiation damage
6. Achieve highest possible completeness and multiplicity and look at $CC_{1/2}$ and CC_{anom} fall-off to judge the data quality and usefulness; try paired refinement
7. Good Luck

References

1. T. C. Redman, *Data Driven: Profiting from Our Most Important Business Asset*, Harvard Business Press, Boston, 2008.
2. http://www.iso.org. Last accessed January 2017.
3. A. Wlodawer, W. Minor, Z. Dauter and M. Jaskolski, *FEBS J.*, 2008, **275**, 1.
4. J. Pflugrath, *Acta Crystallogr., Sect. F*, 2015, **71**, 622.
5. D. Borek, W. Minor and Z. Otwinowski, *Acta Crystallogr., Sect. D*, 2003, **59**, 2031.
6. D. G. Waterman, G. Winter, J. M. Parkhurst, L. Fuentes-Montero, J. Hattne, A. Brewster, N. K. Sauter and G. Evans, *CCP4 Newsl.*, 2013, **49**, 16.
7. A. G. W. Leslie, *Acta Crystallogr., Sect. D*, 1999, **55**, 1696.
8. Z. Otwinowski and W. Minor, *Methods Enzymol.*, 1997, **276**, 307.
9. J. E. Pflugrath, *Acta Crystallogr., Sect. D*, 1999, **55**, 1718.
10. W. Kabsch, *Acta Crystallogr., Sect. D*, 2010, **66**, 125.
11. W. Kabsch, *Acta Crystallogr., Sect. D*, 2010, **66**, 133.
12. P. R. Evans and G. N. Murshudov, *Acta Crystallogr., Sect. D*, 2013, **69**, 1204.
13. M. Mueller, M. Wang and C. Schulze-Briese, *Acta Crystallogr., Sect. D*, 2012, **68**, 42.
14. K. N. Trueblood, H.-B. Buergi, H. Burzlaff, J. D. Dunitz, C. M. Gramaccioli, H. H. Schulz, U. Shmueli and S. C. Abrahams, *Acta Crystallogr., Sect. A*, 1996, **52**, 770.
15. C. Katayama, *Acta Crystallogr., Sect. A*, 1986, **42**, 19.
16. R. H. Blessing, *Acta Crystallogr., Sect. A*, 1995, **51**, 33.
17. R. H. Blessing, *J. Appl. Crystallogr.*, 1997, **30**, 421.
18. E. Garman, *Curr. Opin. Struct. Biol.*, 2003, **13**, 545.
19. P. O'Neill, D. L. Stevens and E. F. Garman, *J. Synchrotron Radiat.*, 2002, **9**, 329.
20. R. B. G. Ravelli and S. M. McSweeney, *Structure*, 2000, **8**, 315.
21. W. P. Burmeister, *Acta Crystallogr., Sect. D*, 2000, **56**, 328.
22. C. Nave and E. F. Garman, *J. Synchrotron Radiat.*, 2005, **12**, 257.
23. J. W. Murray, E. Rudiño-Peñera, R. L. Owen, M. Grininger, R. B. G. Ravelli and E. F. Garman, *J. Synchrotron Radiat.*, 2005, **12**, 268.
24. J. M. Holton, *J. Synchrotron Radiat.*, 2009, **16**, 133.
25. E. F. Garman, *Acta Crystallogr., Sect. D*, 2010, **66**, 339.
26. M. Warkentin, B. J. Hopkins, R. Badeau, A. M. Mulichak, L. J. Keefe and R. E. Thorne, *J. Synchrotron Radiat.*, 2013, **20**, 7.
27. P. R. Evans, *Acta Crystallogr., Sect. D*, 2006, **62**, 72.
28. P. R. Evans, *Acta Crystallogr., Sect. D*, 2011, **67**, 282.
29. K. Diederichs, *Acta Crystallogr., Sect. D*, 2010, **66**, 733.
30. T. R. Schneider and G. M. Sheldrick, *Acta Crystallogr., Sect. D*, 2002, **58**, 1772.
31. U. W. Arndt, R. A. Crowther and J. F. W. Mallett, *J. Phys. E: Sci. Instrum.*, 1968, **1**, 510.
32. A. J. C. Wilson, *Acta Crystallogr.*, 1950, **3**, 397.

33. K. Diederichs and P. A. Karplus, *Nat. Struct. Biol.*, 1997, **4**, 269.
34. M. S. Weiss, *J. Appl. Crystallogr.*, 2001, **34**, 130.
35. T. Krojer and F. von Delft, *J. Synchrotron Radiat.*, 2011, **18**, 387.
36. P. A. Karplus and K. Diederichs, *Science*, 2012, **336**, 1030.
37. K. Diederichs and P. A. Karplus, *Acta Crystallogr., Sect. D*, 2013, **69**, 1215.
38. N. A. Rahman, *A Course in Theoretical Statistics*, Griffin, London, 1968.
39. G. Assmann, W. Brehm and K. Diederichs, *J. Appl. Crystallogr.*, 2016, **49**, 1021.
40. R. L. Owen, E. Rudiño-Piñera and E. F. Garman, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 4912.
41. K. Diederichs, *Acta Crystallogr., Sect. D*, 2006, **62**, 96.
42. G. Winter, PhD thesis, University of Manchester, England, 2009.
43. K. Diederichs, S. McSweeney and R. B. G. Ravelli, *Acta Crystallogr., Sect. D*, 2003, **59**, 903.
44. J. Foadi, P. Aller, Y. Alguel, A. Cameron, D. Axford, R. L. Owen, W. Armour, D. G. Waterman, S. Iwata and G. Evans, *Acta Crystallogr., Sect. D*, 2013, **69**, 1617.
45. K. Diederichs, *Acta Crystallogr., Sect. D*, 2017, **73**, 286.
46. G. N. Murshudov, *Appl. Comput. Math.*, 2011, **10**, 250.
47. M. Strong, M. R. Sawaya, S. Wang, M. Phillips, D. Cascio and D. Eisenberg, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 8060.
48. A. Thorn, J. Parkhurst, P. Emsley, M. Vollmar, R. Nicholls, G. Evans and G. Murshudov, *Acta Crystallogr., Sect. D*, 2017, **73**, 729.
49. S. Brockhauser, R. G. B. Ravelli and A. A. MaCarthy, *Acta Crystallogr., Sect. D*, 2013, **69**, 1241.
50. R. A. Kirian, X. Wang, U. Weierstall, K. E. Schmidt, J. C. H. Spence, M. Hunter, P. Fromme, T. White, H. N. Chapman and J. Holton, *Opt. Express*, 2010, **18**, 5713.
51. J. M. Parkhurst, G. Winter, D. G. Waterman, L. Fuentes-Montero, R. J. Gildea, G. N. Murshudov and G. Evans, *J. Appl. Crystallogr.*, 2016, **49**, 1912.
52. J. M. Parkhurst, A. Thorn, M. Vollmar, G. Winter, D. G. Waterman, L. Fuentes-Montero, R. J. Gildea, G. N. Murshudov and G. Evans, *IUCrJ*, 2017, **4**, 626.
53. Z.-J. Liu, L. Cheng, D. Wu, W. Ding, H. Zhang, W. Zhou, Z.-Q. Fu and B. C. Wang, *Acta Crystallogr., Sect. A*, 2011, **67**, 544.
54. Q. Liu, Z. Zhang and W. A. Hendrickson, *Acta Crystallogr., Sect. D*, 2011, **67**, 45.
55. Q. Liu, T. Dahmane, Z. Zhang, Z. Assur, J. Brasch, L. Shapiro, F. Mancia and W. A. Hendrickson, *Science*, 2012, **336**, 1033.
56. R. Giordano, R. M. F. Leal, G. P. Bourenkov, S. McSweeney and A. N. Popov, *Acta Crystallogr., Sect. D*, 2012, **68**, 649.
57. Q. Liu, Q. Liu and W. A. Hendrickson, *Acta Crystallogr., Sect. D*, 2013, **69**, 1314.
58. D. L. Akey, W. C. Brown, J. R. Konwerski, C. M. Ogata and J. L. Smith, *Acta Crystallogr., Sect. D*, 2014, **70**, 2719.

CHAPTER 6

Structure Determination at Low-resolution, Anisotropic Data and Crystal Twinning

JINDRICH SYMERSKY,[a] YI GUO,[a] JIMIN WANG[b] AND MIN LU*[a]

[a] Rosalind Franklin University of Medicine & Science, Department of Biochemistry & Molecular Biology, 3333 Green Bay Road, North Chicago, IL, USA; [b] Yale University, Department of Molecular Biophysics & Biochemistry, 266 Whitney Avenue, New Haven, CT, USA
*Email: min.lu@rosalindfranklin.edu

6.1 Introduction

Multidrug transporters belonging to the Multidrug And Toxic compound Extrusion (MATE) protein family are integral membrane proteins that can move cytotoxic chemicals across cell membranes.[1] Known MATE transporters utilize either the Na^+ or H^+ electrochemical gradient, and fall into the NorM, DinF and eukaryotic subfamilies.[2] Since MATE proteins can render cells resistant to a wide variety of drugs, they hold great promise as potential therapeutic targets for overcoming multidrug resistance.[3] Despite its medical relevance, the molecular mechanism underlying MATE-mediated multidrug transport is not fully understood. Molecular structures of MATE transporters are essential for elucidating their transport mechanisms. To date, the X-ray structures of four NorM and DinF transporters have been reported, providing insights into how the MATE transporters interact with drug substrates and how they couple the influx of cations (Na^+ or H^+) to the efflux of structurally and chemically distinct drugs.[4–8] Among these reported

Chemical Biology No. 8
Protein Crystallography: Challenges and Practical Solutions
Edited by Konstantinos Beis and Gwyndaf Evans
© The Royal Society of Chemistry 2018
Published by the Royal Society of Chemistry, www.rsc.org

structures is that of a NorM transporter from *Neisseria gonorrhoeae* (NorM-NG), portraying the MATE transporter in an extracellular-facing, substrate-bound state.[5]

The structure determination of NorM-NG had been hindered by severe defects in the protein crystals, which diffracted X-rays both weakly and anisotropically. To overcome these defects, we identified a crystallization chaperone called monobody that bound NorM-NG with nanomolar affinity and improved its crystallization behaviors.[5] We subsequently determined the structure of NorM-NG in complex with the monobody by combining molecular replacement and Multiple-Wavelength Isomorphous Replacement with Anomalous Scattering (MIRAS) phasing. Although the overall structure of NorM-NG bound to the monobody resembles those of other MATE transporters, which were determined in the absence of any crystallization chaperone,[4,6] it remains unclear as to whether the binding of the monobody has altered the structure of NorM-NG.

To address this question, we examined the X-ray diffraction data collected from the NorM-NG crystals, which were obtained in the absence of any added substrate. Our analysis revealed that the NorM-NG crystals were pseudo-merohedrally twinned, with an estimated twin fraction of 0.4–0.5. Despite the challenges posed by the low resolution, data anisotropy and crystal twinning, we determined the NorM-NG structure by using molecular replacement and verified the correctness of the structure solution by conducting difference Fourier analysis. We found that the NorM-NG structure is largely identical to that of NorM-NG in complex with the monobody. Over the course of this work, we have learned a number of lessons in analyzing weak, anisotropic and twinned X-ray diffraction data, which may be valuable for structural biologists who face similar technical challenges when examining crystals with analogous issues.

6.2 Experimental

6.2.1 Protein Expression and Purification

NorM-NG was expressed and purified as follows. Briefly, the pET-15b derived expression vector containing the gene encoding NorM-NG was introduced into *E. coli* BL21 (DE3) cells, which were grown in Luria-Bertani media to an optical density of 0.6 at 600 nm and induced with 1 mM isopropyl β-D-1-thiogalatopyranoside at 37 °C for 3 h.[5] Cells were disrupted by using a microfluidizer at 6 °C. Cell membranes were collected by ultracentrifugation and extracted with 1% n-dodecyl-β-maltoside (DDM) in 20 mM Hepes, pH 7.5, 100 mM NaCl, 20% glycerol and 1 mM tris-2-carboxyethyl-phosphine (TCEP) at 4 °C. The soluble fraction was passed onto a Ni-NTA column pre-equilibrated in 20 mM Hepes, pH 7.5, 100 mM NaCl, 0.05% DDM, 20% glycerol and 1 mM TCEP at 4 °C. Protein was eluted by using the same buffer supplemented with 450 mM imidazole and further purified by using gel filtration chromatography at 4 °C.

6.2.2 Protein Crystallization and Derivatization

NorM-NG was concentrated to ~10 mg ml^{-1} and dialyzed extensively against 20 mM Tris-HCl, pH 8.0, 0.05% DDM, 20% glycerol and 1 mM TCEP at 4 °C. The crystallization of NorM-NG was performed at 22 °C by using the hanging-drop vapor-diffusion method. The protein sample was mixed with equal volume of a crystallization solution containing 100 mM Tris-HCl, pH 9.0, 0.05% DDM and 30% PEG400. For heavy-atom derivatization, the NorM-NG crystals were incubated with 5–10 mM heavy-metal compounds for >2 h at 22 °C.

6.2.3 X-ray Diffraction Data Collection and Analysis

Prior to data collection, the NorM-NG crystals were plunged into liquid nitrogen. X-ray diffraction data were collected from the flash-frozen crystals at beam lines 22-ID and 23-ID in Argonne National Laboratory, using MAR300 CCD detectors, with a sample-to-detector distance of 400 mm. For heavy-metal-treated crystals, fluorescence scans around the metal absorption edges were carried out on the frozen crystals using the automated procedures implemented at the beamlines. X-ray diffraction data were then collected at the wavelengths of the absorption peaks as suggested by the fluorescence scans to maximize the anomalous signals. More than 4000 NorM-NG crystals were screened by using robotic crystal handlers and the resulting X-ray diffraction images were examined by using the software HKL2000 or MOSFLM.[9,10] Further X-ray data analysis was performed using the program suites CCP4 and PHENIX, unless otherwise specified.[11,12] Model building and electron density inspection were conducted by using the software O or COOT.[13,14] Structure refinement was carried out by using the program REFMAC.[15] Figures were prepared with PyMOL.[16]

6.3 Results and Discussion

6.3.1 Data Collection and Radiation Damage

The NorM-NG crystals grew as thin plates, with typical dimensions of 200×50×20 μm. Both streak-seeding and post-crystallization dehydration were attempted vigorously but failed to improve the morphology or diffraction property of the crystals. Furthermore, substantial differences in the diffraction properties were found among the NorM-NG crystals, even from the same batch of the protein sample or the same crystallization drops. Therefore, many NorM-NG crystals had to be screened at the synchrotron beam-lines before a good diffraction quality crystal could be identified and data collected.

Moreover, a high-brilliance synchrotron radiation source equipped with undulator devices was essential for obtaining the best possible diffraction data from the NorM-NG crystals, which diffracted X-rays weakly. The best

NorM-NG crystals diffracted beyond 3.7 Å resolutions, but such ''high-resolution'' diffraction feature was visible only with sufficient X-ray exposure, typically with a 2-fold attenuated beam for 1 s. With such high X-ray dose, the radiation damage suffered by the NorM-NG crystals became an issue.[17] This problem was manifested by the observation that reflections beyond 4.0 Å resolutions disappeared, as shown by visual inspection, after ~30 diffraction images (1° oscillation per image) had been collected. Moreover, the later-collected reflections often had B factors exceeding 20 Å^2 during data reduction, which is consistent with significant radiation damage.

To mitigate the radiation damage and to attain X-ray diffraction data sets as complete as possible, we reduced the X-ray exposure to such a degree that reflections at ~4.0 Å, rather than 3.7 Å resolutions were visible (with beam typically attenuated 3-fold). The NorM-NG crystals were mounted with their longest dimensions aligned perpendicular to the beam and ~30° wedges of data were collected from each spot on the crystal before severe radiation damage occurred. For a typical NorM-NG crystal and a 50×50 μm collimated X-ray beam, after each wedge of data was collected, we shifted the crystal by ~80 μm along the oscillation axis to expose fresh regions to X-rays. We subsequently selected the most isomorphous wedges of diffraction data collected from the same or different crystals, and then merged and scaled the diffraction data together, using the .x files for individual images in HKL2000.[9] On average, each data set examined in this study was derived from 15–20 wedges of data, corresponding to an overall completeness of >80% and redundancy of ~10.

6.3.2 Data Reduction and Anisotropy

Diffraction pattern indexing indicated that the data could accommodate both *P*1 and *C*2 space groups and subsequent merging and scaling yielded similar statistics for the *P*1 and *C*2 space groups (Table 6.1), suggesting that the space group was likely to be *C*2. By contrast, processing the diffraction data in a higher-symmetry space group, including an orthorhombic space group, would lead to a high level of rejected reflections (>90%) and an exceedingly high R_{sym} (>40%). Assuming a solvent/detergent content of 50%, there would be 4 and 2 NorM-NG molecules per asymmetric unit for the *P*1 and *C*2 space groups, respectively.[18] Furthermore, we observed that the NorM-NG crystals diffracted rather anisotropically, *i.e.*, they diffracted X-rays to markedly different Bragg spacings along distinct directions.

For the data processed in the *C*2 space group, they extended to ~3.8 Å resolution along the ***l*** direction, but lower than 5 Å resolution along the ***h*** or ***k*** direction. By using the program PHASER in CCP4, we found that the B factors for data along the best and worst directions differed by >110 Å^2. Such high level of data anisotropy, although undesirable, is not uncommon for crystals grown from detergent-purified solutions, including NorM-NG. It is generally thought that the detergent-covered transmembrane domains in

Table 6.1 Data collection statistics for NorM-NG.

Data	Native	Mercury	Native
Space group	*P*1	*P*1	*C*2
Unit cell dimensions			
a (Å)	73.9	73.3	74.4
b (Å)	89.1	88.4	163.6
c (Å)	107.9	107.6	108.7
α (°)	84.1	83.2	90.0
β (°)	75.3	74.9	104.9
γ (°)	65.6	65.5	90.0
Resolution (Å)	50–3.8 (3.87–3.8)	50–4.0 (4.07–4.0)	50–3.8 (3.87–3.8)
R_{sym}	0.097 (0.487)	0.119 (0.751)	0.100 (0.448)
R_{pim}	0.032 (0.171)	0.032 (0.153)	0.024 (0.133)
R_{meas}	0.118 (0.459)	0.110 (0.454)	0.118 (0.471)
$\langle I/\sigma I \rangle$	18.0 (1.2)	18.8 (1.8)	21.4 (1.7)
Completeness (%)	91.8 (65.8)	86.9 (51.3)	94.8 (66.1)
Redundancy	11.4 (5.3)	13.1 (9.3)	19.8 (8.4)
No. of unique/total reflections	22 113/250 996	17 827/232 588	11 782/232 744
Wilson B factors (Å^2)	180	140	140
Twinning operator	$-\boldsymbol{h}, -\boldsymbol{h}+\boldsymbol{k}, -\boldsymbol{l}$	$-\boldsymbol{h}, -\boldsymbol{h}+\boldsymbol{k}, -\boldsymbol{l}$	N/A
Twin fraction from Britton plot	0.35	0.39	N/A

integral membrane proteins are incapable of forming strong crystal-packing interactions between neighboring protein molecules, as opposed to the extracellular and/or intracellular protein loops, which can make stronger and more specific crystal-packing interactions.[19]

Although not unexpected, we reasoned that the data anisotropy seen in the X-ray diffraction would pose several problems for further data analysis. Specifically, it has been demonstrated that data anisotropy can obscure the detection of pathological defects of protein crystals, including twinning.[20] Moreover, for the calculation and inspection of experimentally-phased electron density maps, data anisotropy may exert negative effects, since the diffraction data are rather noisy and of low resolution along the poorly diffracting directions. Based on such considerations, we initiated our data analysis with molecular replacement, since the structure of NorM-NG bound to monobody had been determined, and since the commonly used software for molecular replacement can account for data anisotropy.

6.3.3 Twinning Detection and Analysis

We used the structure of monobody-bound NorM-NG as the search model for molecular replacement using the program MOLREP or PHASER and found only a one-molecule solution for data processed in the *C*2 space group. This was unexpected since the self-rotation function (Figure 6.1) and packing considerations both suggested that more than one NorM-NG molecules are present in the asymmetric unit. In addition, structure refinement

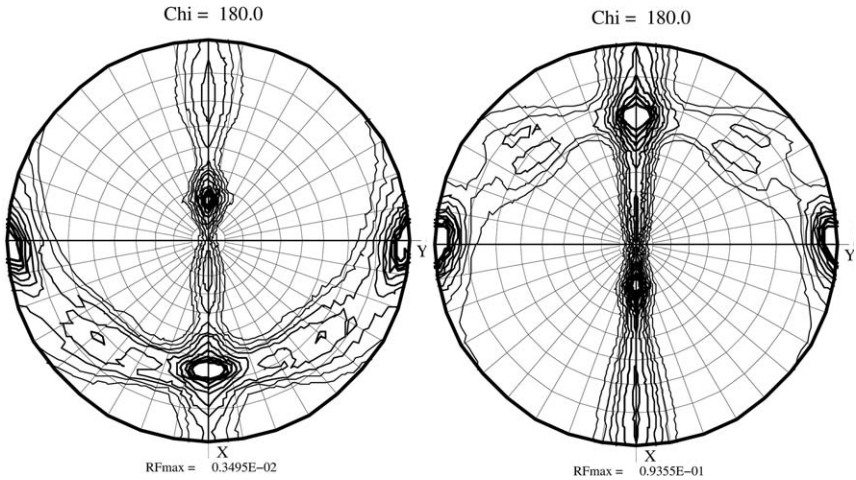


Figure 6.1 Self-rotation function for NorM-NG. The self-rotation function was calculated using data from 50–4.5 Å resolutions and an integration radius of 37 Å. Data processed in the *P*1 space group was used in the left panel and that in the *C*2 space group in the right panel. Peaks on this section ($\chi = 180$) suggest the positions of three two-fold symmetry axes in both the *P*1 and *C*2 space groups.

in the *C*2 space group stalled at an R_{free} of >43%. Furthermore, we failed to verify the correctness of this structure solution by conducting difference Fourier analysis on heavy-atom derivatives (see below). We also noticed that the intensity statistics for X-ray data processed in the *C*2 space group suggested that the data are substantially different from those expected for untwinned data, although no twin law could be found in this space group.

To examine the possibility of crystal twinning, we analyzed the data processed in the *P*1 space group by using the *L*-test, which has been developed to overcome the difficulties in detecting crystal twinning in anisotropic data.[20] As shown in Figure 6.2, the results from the *L*-test suggested that the NorM-NG data are substantially twinned, and the analysis based on the Britton plot indicated that the diffraction data has a twin fraction of 0.35.[21] Furthermore, we identified a twinning operator of $(-\boldsymbol{h}, -\boldsymbol{h}+\boldsymbol{k}, -\boldsymbol{l})$, by using XTRIAGE in PHENIX or CTRUNCATE in CCP4. As noted above, the *L*-test for data processed in the *C*2 space group also suggested that the data are twinned, implying that the detected twinning is unlikely caused by treating crystallographic symmetry as non-crystallographic symmetry (NCS), *i.e.*, processing data conforming to the *C*2 symmetry in a reduced symmetry space group *P*1.

Therefore, the NorM-NG crystals seemed to suffer from pseudo-merohedral twinning,[22] probably arising from fortuitous unit cell geometry (see Figure 6.3). The *P*1 space group (Table 6.1) appears to represent the ''true'' crystal symmetry, whereas the *C*2 space group seems to correspond to the higher symmetry caused by crystal twinning. It is apparent that the axis

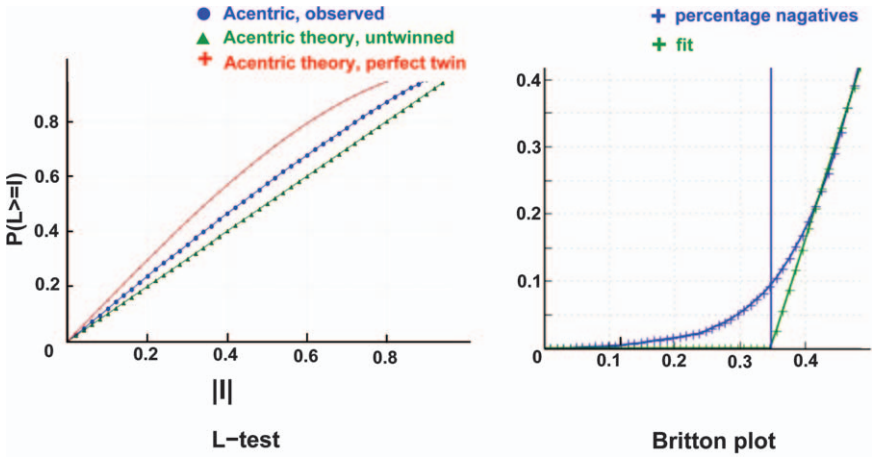


Figure 6.2 *L*-test of diffraction data and twin fraction estimation. Left panel: *L*-test for twinning was conducted using the native data processed in the *P*1 space group. Theoretical distribution of acentric data is shown by a green curve, that for perfectly twinned (*i.e.*, twin fraction = 0.5) data corresponds to a red curve. Right panel: Britton plot showing the fraction of negative intensities after detwinning as a function of the assumed twin fraction. Blue line highlights the estimated twin fraction.

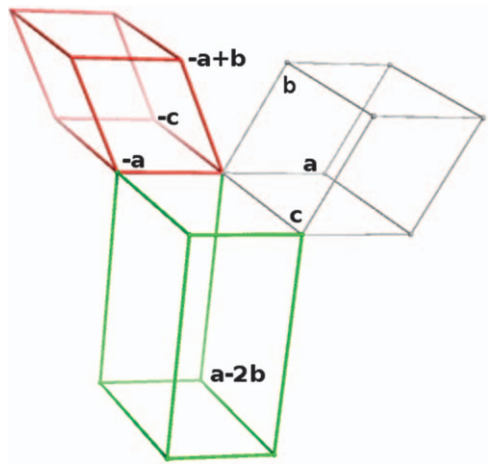


Figure 6.3 Unit cell relationship. The *P*1 cell is drawn in gray, unit cell vectors are $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$, respectively. The twin-related *P*1 cell is in red, its unit cell vectors are $-\boldsymbol{a}$, $-\boldsymbol{a}+\boldsymbol{b}$, and $-\boldsymbol{c}$, respectively. The *C*2 cell is in green, its unit cell vectors are $-\boldsymbol{a}$, $\boldsymbol{a}-2\boldsymbol{b}$, and $\boldsymbol{c}$, respectively.

of the twinning operator $(-\boldsymbol{h}, -\boldsymbol{h}+\boldsymbol{k}, -\boldsymbol{l})$ coincides with the two-fold symmetry axis in the *C*2 unit cell (Figure 6.3), thereby explaining the monoclinic symmetry found during data merging and scaling (Table 6.1). Furthermore, as previous studies have shown, twin fraction can vary considerably among

crystals and even within the same crystal.[22] However, we had not found any substantial differences in estimated twin fractions in data collected from different NorM-NG crystals, provided that they were analyzed by using the same software and resolution cutoffs, *etc.*

This result implied that we could not overcome the twinning problem simply by selecting data collected from untwinned crystals or crystals with negligible twinning (*i.e.*, with twin fractions <0.2). In addition, to test whether the detected twinning was introduced by merging data from different crystals, we further attenuated the beam (typically by 9-fold instead of 3-fold) and collected ''large'' wedges of diffraction data (180° instead of 30°) from NorM-NG crystals. These large wedges were individually processed in the *P*1 space group to ~4 Å and >80% complete, which all behaved the same way in the twinning analysis (*e.g.*, *L*-test) as those data sets derived from multiple ''small'' wedges. This finding strongly suggested that the detected twinning is intrinsic to the NorM-NG crystals, thus ruling out the possibility that the twinning was caused inadvertently by the merging of multiple wedges of diffraction data.

6.3.4 Molecular Replacement and Structure Solution

Taking the crystal twinning into consideration, we pursued molecular replacement further by using the data processed in the *P*1 space group. We expected to find a correct molecular replacement solution in the true unit cell. If so, subsequent structure refinement will further verify the crystal twinning. We examined both the uncorrected and twin-corrected data, and by using the uncorrected data, we found a two-molecule solution in the *P*1 space group, with an estimated solvent/detergent content of ~80% and a reasonable crystal packing arrangement (see below). Those two NorM-NG molecules were related by a two-fold screw NCS axis in the asymmetric unit (Figure 6.4), which is perpendicular to the axis corresponding to the twinning operator $(-\boldsymbol{h}, -\boldsymbol{h}+\boldsymbol{k}, -\boldsymbol{l})$. This packing arrangement is consistent with the self-rotation function (Figure 6.1), in contrast to the one-molecule solution found in the *C*2 space group.

Structure refinement of the two-molecule solution to 3.8 Å resolution yielded an R_{free} of ~38% and a structure model with reasonable stereochemistry. Further phase improvement *via* density modification, including solvent flattening and two-fold NCS averaging, was subsequently performed. Despite the high solvent content of NorM-NG crystals, solvent flattening only improved the electron density map slightly. This seemingly unexpected result may be due to the facts that our data are weak and of modest resolution, and the unit cell is relatively small, which all limit the effectiveness of solvent flattening.[23,24] The density-modified electron density map was calculated to 3.8 Å resolution and used for subsequent model-rebuilding. Final results of the structure refinement, with or without using the twinning operator, are summarized in Table 6.2. In both cases, two-fold NCS restraints were used throughout the structure refinement in order to improve the

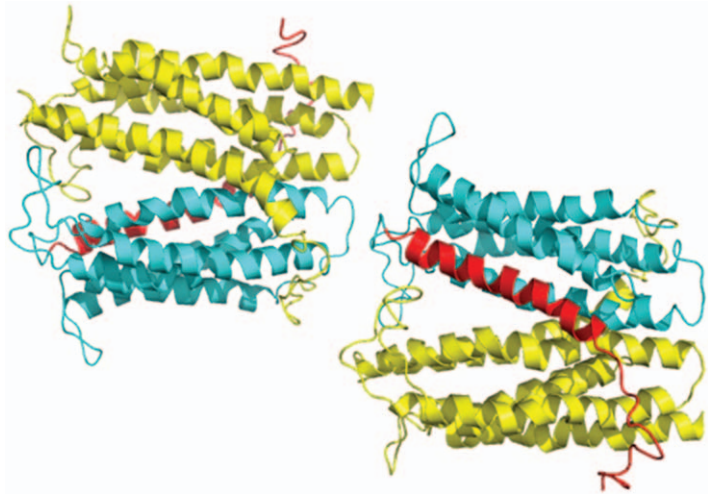


Figure 6.4 The two-molecule solution found in the *P*1 space group. The N- and C-terminal domains of NorM-NG are colored cyan and yellow, respectively, except for the first transmembrane helices (TM1), which are colored in red.

Table 6.2 Refinement statistics for NorM-NG.

Data	Native	Native
Space group	*P*1	*P*1
Twinning operator	$-\boldsymbol{h}, -\boldsymbol{h}+\boldsymbol{k}, -\boldsymbol{l}$	None
Twin fraction from REFMAC	0.50	N/A
Resolution (Å)	20–3.8	20–3.8
No. of reflections	20 899	20 899
R/R_{free}	0.328 (0.395)/0.364 (0.440)	0.435 (0.527)/0.482 (0.548)
No. of protein atoms	6852	6852
Average temperature factors (Å^2)	199	175
Rms bond lengths (Å)	0.01	0.02
Rms bond angles (°)	2.0	2.4
Ramachandran plot: allowed (%)	100.0	98.9
Disallowed (%)	0	1.1

observable-to-parameter ratio, although the NCS averaging failed to yield significant improvement to the electron density map.

As shown in Table 6.2, structure refinement with the twinning operator yielded an R_{free} that was 11% lower and a protein model with better stereochemistry, as opposed to that without the twinning operator, supporting that the NorM-NG crystals were substantially twinned. The twin fraction was refined to 0.5, which is consistent with the low R_{sym} obtained for the data processed in the *C*2 space group (Table 6.1). This result also implied

that detwinning of the NorM-NG data would be impossible without introducing a substantial amount of additional noise into the data. For instance, detwinning with a twin fraction 0.48 would increase the estimated intensity error by ~25-fold.[25] Our results also indicated that the estimation of twin fraction based on intensity statistics, such as the Britton plot, underestimated the level of twinning for the NorM-NG crystals (Figure 6.2). This underestimation may result from the presence of NCS coupled with twinning, measurement errors, and/or data anisotropy.[26]

Furthermore, structure refinement with anisotropically corrected diffraction data by using the software PHASER gave rise to structure models with markedly higher R_{free} and worse stereochemistry, as opposed to the uncorrected data. Moreover, electron density maps calculated with the ''corrected'' data were no better than those with the uncorrected data. This result implied that the re-scaling of weak, anisotropic and twinned diffraction data probably introduced additional errors and/or improperly treated the twin-related reflections, as reflected in the markedly increased R_{twin} after the anisotropic ''correction''.[27] In parallel, we also attempted to ellipsoidally truncate the anisotropic data by using the UCLA server, but the truncated data only worsened the structure refinement.[28] Given the modest data-parameter ratio in this case, it seems likely that the ellipsoidal truncation of the diffraction data reduced the data completeness to such a degree that it offset the benefit of removing weak and non-existent reflections.

Based on the refined model, we conducted the RvR analysis using XTRIAGE in PHENIX to gain further insights into crystal twinning and space group assignment.[27] This analysis revealed that with respect to the twinning operator $(-\boldsymbol{h}, -\boldsymbol{h}+\boldsymbol{k}, -\boldsymbol{l})$, R_{twin}^{obs} (0.08–0.27) is consistently lower than R_{twin}^{calc} (0.46–0.52) throughout the resolution range (50–3.8 Å), indicating that severe twinning is indeed associated with this twinning operator and unlikely due to a rotational pseudo-symmetry. Moreover, the *P*1 space group is unlikely to be mis-specified, since in the case of a mis-specified crystal symmetry, both R_{twin}^{obs} and R_{twin}^{calc} would be close to zero.[27,29] Taken together, our RvR analysis supported the *P*1 space group assignment and verified the twinning operator $(-\boldsymbol{h}, -\boldsymbol{h}+\boldsymbol{k}, -\boldsymbol{l})$.

6.3.5 Difference Fourier Analysis

Given the modest quality of our X-ray data, we sought to verify the correctness of the structure solution by conducting an analysis that is independent from structure refinement. To this end, we incubated the NorM-NG crystals with >20 different heavy-metal compounds. We then employed the inverse-beam method to collect X-ray diffraction data near the heavy-metal absorption edges of these crystals, aiming to measure the diffraction intensities of Friedel pairs as accurately as possible. We subsequently used the model-derived or density-modified phases to carry out difference anomalous Fourier analysis. However, we were unable to identify any strong difference

anomalous Fourier peaks that are located near reactive amino acids, for instance, Cys residues for mercury compounds.

In parallel, we conducted difference isomorphous Fourier analysis. For a mercury (thimerosal) derivative, we found four distinct peaks per asymmetric unit at contour levels higher than 7σ in the difference Fourier maps calculated using the model-derived phases (Figure 6.5). The same maps did not reveal any other peaks greater than 4σ. Among the four strong peaks, two were located in close proximity to the Cys381 residues from the two NorM-NG molecules, whereas the other two peaks were near residues Ala264, Ile287 and Ser290, making little chemical sense. However, after the fractional coordinates of the two twin-related NorM-NG molecules were generated by applying the transpose of the inverse of the matrix corresponding to the twinning operator, the two ''inexplicable'' difference Fourier peaks were placed into close proximity to the two Cys381 residues within the twin-related NorM-NG molecules (Figure 6.5).

A plausible explanation for the twin-related difference Fourier peaks is that the size of the twin domains in the NorM-NG crystals is similar to the X-ray beam coherence length, which enabled the X-ray waves scattered from the two domains to interfere.[22] As such, the observed intensities are derived from the complex average of the scattering factors from each twin domain and it is thus conceivable that the difference isomorphous Fourier maps showed features for mercury binding in both twin domains. In accord with this scenario, we observed several, but not all, cross-peaks between the mercury-binding sites within the two twin domains in the difference isomorphous Patterson maps (see below), implying that the size of the twin domains in the NorM-NG crystals approaches the beam coherence length.[22]

Similar observations in difference isomorphous Fourier analysis were made by others, although the presence of additional difference Fourier peaks were attributed to the similar molecular packing between the twin

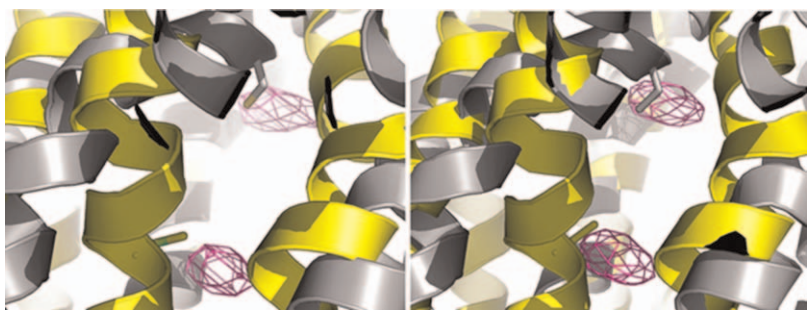


Figure 6.5 Difference isomorphous Fourier maps. The twin-related mercury-binding sites in the two NorM-NG molecules (yellow) in the asymmetric unit are highlighted by the difference Fourier peaks (magenta mesh, 7σ contour level). The twin-related NorM-NG molecules are colored in gray. Residues Cys381 are drawn as stick models. Data from 15 to 3.8 Å resolution were used for map calculation.

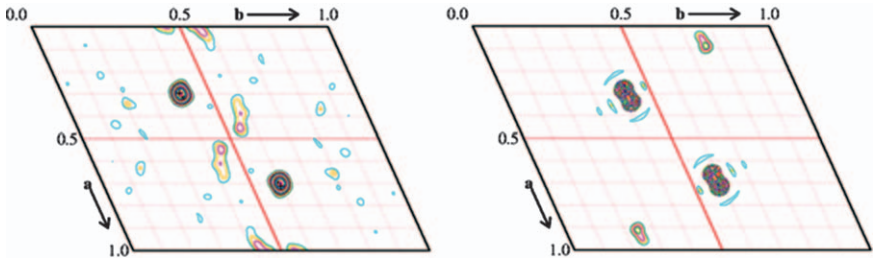


Figure 6.6 Difference isomorphous Patterson maps. Left panel: observed difference Patterson maps for the mercury derivative at $w = 0.5$, data were processed in the *P*1 space group. The peaks (16.6 σ) are the superposition of the cross-peaks between the two mercury atoms within the two twin domains. Cross-peaks in one twin domain are highlighted by "+" and those in the second domain are marked by "*". Right panel: predicted difference Patterson based on the coordinates of the mercury-binding sites. This figure was prepared using XPREP (Bruker); data from 15 to 4 Å resolutions were used for this analysis.

domains.[30] Furthermore, besides crystal twinning, static disorder can also give rise to additional, "inexplicable" difference Fourier peaks.[31] Static disorder in principle can be distinguished from crystal twinning based on the diffraction intensity statistics, because these two defects affect crystal diffraction rather differently. As described above, since the results from the *L*-test and RvR plot are both suggestive of crystal twinning in our case, we argue that the NorM-NG crystals suffered from pseudo-merohedral twinning rather than static disorder.

Significantly, the binding of the mercuric compound to Cys381 in NorM-NG, which had been observed in the monobody-bound crystal form, can also explain the difference isomorphous Patterson maps calculated for the mercury derivative.[5] In these maps, we found not only the peaks corresponding to the mercury-binding sites within the same twin domains (Figure 6.6), but also some of the cross-peaks for the binding sites between the twin domains, which are consistent with our difference isomorphous Fourier maps (Figure 6.5). Taken together, the identification of the mercury-binding sites in NorM-NG using model-derived phases verified the correctness of our structure solution and further confirmed the crystal twinning in the *P*1 space group. By contrast, we were unable to identify any heavy-metal binding site by using difference Fourier analysis in the *C*2 space group, suggesting that the correct space group is *P*1, not *C*2.

6.3.6 Attempts in Experimental Phasing

Because unbiased experimental phases would be useful for uncovering potential structural differences between NorM-NG and its monobody-bound form, we set out to calculate SIR phases for the mercury derivative, using either the uncorrected or detwinned diffraction data. We detwinned both the

native and derivative data by using the program DETWIN in CCP4 and used various twin fractions, ranging from 0.1 to 0.4, with an increment of 0.1. Although the peak heights for the twin-related peaks gradually decreased in the difference isomorphous Fourier maps as we increased the twin fractions to 0.4, structure refinement against the ''detwinned'' data always yielded twin fractions greater than 0.4, probably reflecting the difficulty in correcting the twinned data computationally.

Furthermore, the refinement of heavy-atom positions and occupancies using the program SHARP fared better using uncorrected diffraction data than the computationally ''detwinned'' data, which are likely to be noisier than the uncorrected data.[32] The solvent-flattened SIR maps calculated to 6 Å resolution without the two-fold NCS averaging revealed a clear protein-solvent boundary and no additional protein molecules in the asymmetric unit other than the two known molecules (Figure 6.4). The quality of the experimentally phased electron density maps, however, was insufficient for model-building even after extensive density modification including solvent flattening and NCS averaging. Inclusion of the experimental phases also failed to improve the structure refinement. As we only found one heavy-metal derivative, we did not pursue experimental phasing further, since it has been suggested that at least four different heavy-metal derivatives are necessary for MIR phasing using highly twinned X-ray data.[33]

6.3.7 Comparison of the NorM-NG Structures

The final structure model includes residues 8–455. Within the crystals, the neighboring NorM-NG molecules are packed against each other in a 'head-to-tail' fashion *via* interactions between the intracellular- and extracellular-loops, particularly along the ***c*** direction (Figure 6.7). By contrast, the largely hydrophobic surface of the transmembrane domain in NorM-NG, presumably covered by detergents and/or lipids, plays little role in forming the crystal packing interactions.

This crystal packing arrangement helps to explain why the diffraction of the NorM-NG crystals extended to higher resolutions along the ***l*** direction, as opposed to the ***h*** or ***k*** direction, since the neighboring NorM-NG molecules form more extensive packing interactions along the ***c*** direction, which approximately corresponds to the ***l*** direction in reciprocal space. Furthermore, a similar crystal packing arrangement has been observed for the NorM transporter from *V. cholerae*, NorM-VC, which was also crystallized in a detergent-purified form.[4] However, the NorM-VC molecules were packed in a 'tail-to-tail' or 'head-to-head' manner in the crystals, rather than a 'head-to-tail' fashion as seen in NorM-NG.

As observed in the monobody-bound crystal form, NorM-NG adopts an overall V-shaped, extracellular-facing conformation.[5] The C-terminal residues 456–463 in NorM-NG, which were found to make close contacts with the monobody, were unresolved in the current structure and likely to be disordered. The two NorM-NG molecules in the asymmetric unit have largely

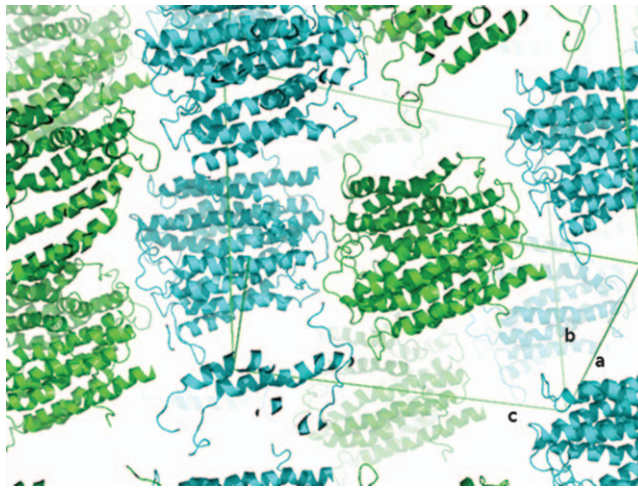


Figure 6.7 Crystal packing of NorM-NG. One molecule in the asymmetric unit is colored in green and the other in cyan. The unit cell axes in the *P*1 space group are highlighted by "***a***", "***b***" and "***c***".

the same structure, with an rms deviation of ~0.3 Å for 448 Cα positions. The structure of NorM-NG is also similar to that of the monobody-bound form, with an rms deviation ~0.6 Å for 448 Cα positions. Therefore, although the NorM-NG has only been refined to 3.8 Å resolution, we argue that the monobody has not altered the transporter structure to any significant extent. Indeed, composite omit maps revealed no major discrepancy between the refined NorM-NG model and electron density.

6.3.8 Ligand-binding Site

In the structure of monobody-bound NorM-NG, we observed a conspicuous electron density feature in the substrate-binding site, which was attributed to an unidentified ligand.[5] No such density feature, however, was found in the electron density maps calculated for NorM-NG. Despite this difference, the two extracellular protein loops that cap the substrate-binding cavity have not altered their conformations in NorM-NG, as compared with their counterparts in the monobody-bound structure. Our data thus raised the question of why NorM-NG remains in its substrate-bound state in the absence of any bound ligand.

The answer to this question may be found in at least three, mutually non-exclusive factors. Firstly, the crystallization conditions for NorM-NG and its monobody-bound complex are different. It is plausible that under the conditions that NorM-NG was crystallized, the binding of the unidentified ligand was suppressed to such a level that it was no longer crystallographically detectable owing to low occupancy. Secondly, the diffraction data for NorM-NG is much weaker, more anisotropic, and of lower resolution, as compared

with that for the monobody-bound complex. Furthermore, well-determined experimental phases were used in calculating the electron density maps for the monobody-bound NorM-NG.[5] As such, the electron density for the unidentified ligand in NorM-NG might be undetectable simply because of the limited quality of the diffraction data, *e.g.*, modest resolution and/or phase bias from the protein-only model. Thirdly, since the extracellular loops that cap the substrate-binding site are intimately involved in the crystal-packing interactions (Figure 6.7), it seems likely that the ligand-bound conformation of NorM-NG was stabilized by the crystal-packing interactions.

6.4 Conclusion

The structure determination of NorM-NG presents a particularly difficult case, because the crystals diffracted X-rays weakly, anisotropically and suffered from severe twinning defects. Despite such issues, we were able to obtain a structure solution thanks to the availability of a very similar structure model. During the course of this study, we have also attempted to use other protein models for molecular replacement, including the published NorM-VC structure, but without success.[4] This failure may be in part due to the amino-acid sequence and structural differences between NorM-NG and NorM-VC, which were reflected in ~33% sequence identity and an rms deviation of >5 Å for 440 common Cα positions, or >3 Å for 300 Cα positions if only the membrane-spanning helices were considered. Indeed, previous studies have suggested that molecular replacement would become increasingly difficult once the rms deviation between search model and target structure exceeds 2 Å.[34]

Another challenge in this structure determination stemmed from the fact that the diffraction data were available only to moderate resolution. At ~4 Å resolution, the ratio of the observables to the parameters is so low that structure refinement becomes less powerful in improving the protein model and it is easy to fall victim to the model bias during map interpretation. Such problems can be mitigated or rectified if high-number NCS restraints and/or accurate experimental phases are available, as they can be included during structure refinement to enhance the observable to parameter ratio.[35] Furthermore, unbiased experimental maps are extremely useful for revealing the subtle differences between the search models and target structures at moderate resolutions, especially if molecular replacement is initially used to solve the phase problem, *i.e.*, to overcome phase bias.[7]

In the case of NorM-NG, we only had two-fold NCS and poor SIR phases, which had made modest and no contribution in improving the structure refinement, respectively. As such, it was essential for us to validate the structure solution by means of difference Fourier analysis. Despite the fact that the NorM-NG crystals were highly twinned, we were able to use model-derived phases to identify a mercury derivative and locate the heavy-atom binding sites. Although the quality of the resulting SIR phases was insufficient for model-building or improving structure refinement, we could

explain both the difference isomorphous Fourier and Patterson maps. Therefore, in situations where only low-resolution, anisotropic and twinned X-ray diffraction data are attainable, we argue that combining molecular replacement with difference Fourier analysis may still lead to a correct structure solution, provided that a similar structure model and a heavy-metal derivative are both available.

Acknowledgements

We thank the beam-line staff at GM/CA-CAT (23-ID) and SER-CAT (22-ID) of Argonne National Laboratory for assistance during X-ray data collection. We also thank T. Yeates (UCLA) and A. Leslie (MRC) for discussions and our anonymous reviewers for their insightful suggestions. This work was supported in part by the National Institutes of Health (R01-GM094195 to Min Lu).

References

1. T. Kuroda and T. Tsuchiya, *Biochim. Biophys. Acta*, 2009, **1794**, 763.
2. M. H. Brown, I. T. Paulsen and R. A. Skurray, *Mol. Microbiol.*, 1999, **31**, 394.
3. M. A. Fischbach and C. T. Walsh, *Science*, 2009, **325**, 1089.
4. X. He, P. Szewczyk, A. Karyakin, M. Evin, W.-X. Hong, Q. Zhang and G. Chang, *Nature*, 2010, **467**, 991.
5. M. Lu, J. Symersky, M. Radchenko, A. Koide, Y. Guo, R. Nie and S. Koide, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 2099.
6. Y. Tanaka, C. J. Hipolito, A. D. Maturana, K. Ito, T. Kuroda, T. Higuchi, T. Katoh, H. E. Kato, M. Hattori, K. Kumazaki, T. Tsukazaki, R. Ishitani, H. Suga and O. Nureki, *Nature*, 2013, **496**, 247.
7. M. Lu, M. Radchenko, J. Symersky, R. Nie and Y. Guo, *Nat. Struct. Mol. Biol.*, 2013, **11**, 1310.
8. M. Radchenko, J. Symersky, R. Nie and M. Lu, *Nat. Commun.*, 2015, **6**, 7995.
9. Z. Otwinowski and W. Minor, *Methods Enzymol.*, 1997, **276**, 307.
10. A. W. G. Leslie, H. R. Powell, G. Winter, O. Swenson, D. Spruce, S. McSweeney, D. Love, S. Kinder, E. Duke and C. Nave, *Acta Crystallogr.*, 2002, **D58**, 1924.
11. M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potteron, H. R. Powell, R. J. Read, A. Vagin and K. S. Wilson, *Acta Crystallogr., Sect. D*, 2011, **67**, 235.
12. P. D. Adams, P. V. Afonie, G. Bunkoczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L. W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, J. S. Richardson, T. C. Terwilliger and P. H. Zwart, *Acta Crystallogr., Sect. D*, 2010, **66**, 213.

13. T. A. Jones, J. Y. Zou, S. W. Cowan and M. Kjeldgaard, *Acta Crystallogr., Sect. A*, 1991, **47**, 110.
14. P. Emsley and K. Cowtan, *Acta Crystallogr., Sect. D*, 2004, **60**, 2126.
15. G. N. Murshudov, A. A. Vagin and E. J. Dodson, *Acta Crystallogr., Sect. D*, 1997, **53**, 240.
16. W. L. DeLano, *The PyMOL Molecular Graphics System*, DeLano Scientific, 2002.
17. P. O'Neill, D. L. Stevens and E. F. Garman, *J. Synchrotron Radiat.*, 2002, **9**, 329.
18. B. W. Matthews, *J. Mol. Biol.*, 1968, **33**, 491.
19. H. Michel, in *Crystallization of Membrane Proteins*, ed. H. Michel, CRC Press, 1991, pp. 73–88.
20. J. E. Padilla and T. O. Yeates, *Acta Crystallogr., Sect. D*, 2003, **59**, 1124.
21. D. Britton, *Acta Crystallogr., Sect. A*, 1972, **28**, 296.
22. T. O. Yeates, *Methods Enzymol.*, 1997, **276**, 344.
23. B.-C. Wang, *Methods Enzymol.*, 1985, **115**, 90.
24. J. P. Abrahams and N. Ban, *Methods Enzymol.*, 2003, **374**, 163.
25. Z. Dauter, *Acta Crystallogr., Sect. D*, 2003, **59**, 2004.
26. J. G. Wittmann and M. G. Rudolph, *Acta Crystallogr., Sect. D*, 2007, **63**, 744.
27. A. A. Lebedev, A. A. Vagin and G. N. Murshudov, *Acta Crystallogr., Sect. D*, 2006, **62**, 83.
28. M. Strong, M. R. Sawaya, S. Wang, M. Phillips, D. Cascio and D. Eisenberg, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 8060.
29. T. Barends, R. deJong, K. van Straaten, A.-M. Thunnisen and B. Dijkstra, *Acta Crystallogr., Sect. D*, 2005, **61**, 613.
30. Y.-R. Yuan, O. Martsinkevich and J. F. Hunt, *Acta Crystallogr., Sect. D*, 2003, **59**, 225.
31. S. A. Shah and A. T. Brunger, *J. Mol. Biol.*, 1999, **285**, 1577.
32. E. de La Fortelle and G. Bricogne, *Methods Enzymol.*, 1997, **276**, 472.
33. T. O. Yeates and D. Rees, *Acta Crystallogr., Sect. A*, 1987, **43**, 30.
34. T. C. Terwilliger, R. J. Read, P. D. Adams, A. T. Brunger, P. V. Afonie, R. W. Grosse-Kunstleve and L.-W. Hung, *Acta Crystallogr., Sect. D*, 2012, **68**, 861.
35. A. T. Brunger, B. DeLaBarre, J. M. Davies and W. I. Weis, *Acta Crystallogr., Sect. D*, 2009, **65**, 128.

CHAPTER 7

Structure Determination and Refinement of Large Macromolecular Assemblies at Low Resolution

RICHARD D. BUNKER

Friedrich Miescher Institute for Biomedical Research, Maulbeerstasse 66, Basel 4058, Switzerland
Email: richard.bunker@fmi.ch

7.1 Introduction

The crystal structure of COP9 signalosome (CSN) was originally reported by Thomä and coworkers,[1] with an extended description of the crystallographic methods appearing subsequently.[2] A structure-function study of CSN-substrate complexes following on from this crystallographic study was reported more recently.[3] Here, an expanded account of the crystallographic structure determination of CSN is given, covering recent developments.

CSN is a multi-protein complex in eukaryotes with isopeptidase activity that is an essential regulator of intracellular protein degradation. CSN controls the large family of cullin-RING ubiquitin E3 ligases (CRLs) and has important roles in many cellular processes involving the ubiquitin proteasome system for protein turnover; notably DNA repair, cell cycle control, and differentiation.

Chemical Biology No. 8
Protein Crystallography: Challenges and Practical Solutions
Edited by Konstantinos Beis and Gwyndaf Evans
© The Royal Society of Chemistry 2018
Published by the Royal Society of Chemistry, www.rsc.org

Together with the 19S proteasome lid and eukaryotic initiation factor 3, other major regulatory complexes that also control aspects of protein lifespan, CSN defines the PCI (proteasome lid, CSN, eukaryotic initiation factor 3) class of evolutionarily related protein complexes. Human CSN is a complex of eight different proteins, designated CSN1 to CSN8 by decreasing size, with a total molecular weight of ~350 kDa (Figure 7.1). Two CSN

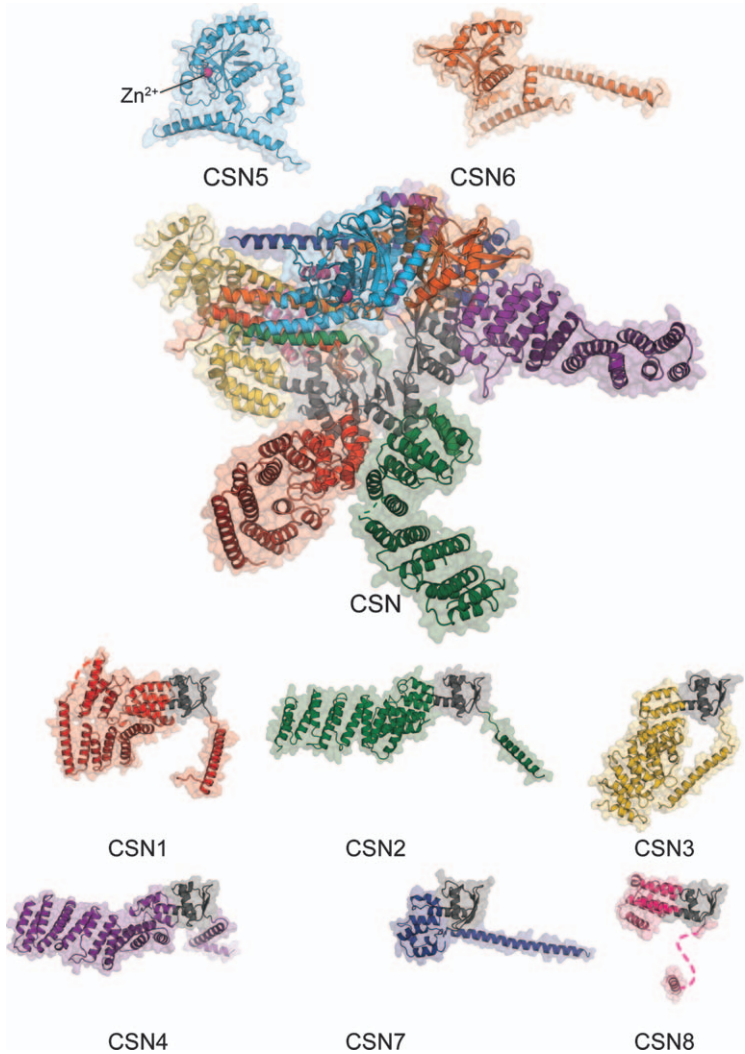


Figure 7.1 Cartoon representations of the COP9 signalosome complex (centre) and its constituents: the two MPN domain protein subunits (CSN5–CSN6) (above) and six PCI domain proteins (CSN1–CSN4 and CSN7a–CSN8) (below). The winged-helix subdomain of each of PCI protein is coloured grey and the Zn^{2+} ion in the active site of CSN5 is labelled. Major loops that were unmodelled because of disorder are indicated by dashed lines. Reprinted under a Creative Commons License (https://creativecommons.org/licenses/by/2.0) from Bunker.[2]

subunits, CSN5 and CSN6, are MPN (MPR1/PAD1 amino-terminal) domain proteins and the other six subunits are PCI domain proteins. MPN domains have a mixed α/β metalloprotease fold, and PCI domain proteins have a bipartite fold consisting of an array of irregular α-helical repeats followed by a winged-helix subdomain. All CSN proteins bear C-terminal α-helical extensions.

The CSN5 and CSN6 MPN domains form a heterodimer, and the six PCI proteins form a horseshoe-shaped ring (PCI ring) with a central intermolecular β-sheet, from which their N-terminal α-helical domains radiate prominently. All eight proteins are united in an intricate bundle formed by their C-terminal α-helices, which stacks between the CSN5–CSN6 heterodimer and PCI ring.

At the outset of our work on the human CSN, limited structural information was available for the complex, its components or the other PCI complexes; the subunit composition and fold classes were established, and CSN5 was known to be catalytic and bind a single Zn^{2+} ion.

Crystal structures for parts of four CSN proteins were available by the end of the project: *Arabidopsis thaliana* CSN1 (residues 32–349 of 441, PDB entry 4LCT[4]); *Homo sapiens* CSN5 MPN domain (residues 2–257 of 334, 4F7O[5]); the MPN domain core of *Drosophila melanogaster* CSN6 (residues 51–184 of 341, 4E0Q[6]); and *A. thaliana* CSN7 (residues 4–164 of 260, 3CHM[7]). Crystal structures were described for paralogues from the 19S proteasome lid (RPN proteins) and initiation factor 3 (elF3 proteins): CSN2 (*D. melanogaster* RPN6, PDB entry 3TXN[8]); and CSN8 (human eIF3k, 1RZ4; and *Schizosaccharomyces pombe* RPN12, 4B0Z[9]). No crystal structures were described for CSN3 or CSN4.

An existing low-resolution electron microscopy (EM) structure of CSN in negative stain (EMDB entry 1700) proved fundamentally different to the crystal structure and unhelpful. A CSN-CRL complex structure also determined by cryo-EM at low resolution and in negative stain (EMDB entry 2173), reported when the crystallography was essentially complete, provided insight into the general interaction of CSN and a substrate.

7.2 Crystallization, Data Collection and Processing

Micro-crystals of full-length human CSN were obtained by initial vapour diffusion screening. The crystallization conditions were 12% (w/v) PEG 6000, 0.1 M trisodium citrate pH 5.4, 0.1 M lithium or ammonium sulfate, with CSN added at a concentration of 10 $mg\,mL^{-1}$ in a solution containing 50 mM HEPES pH 7.4, 200 mM NaCl, 2 mM dithiothreitol or 0.5 mM tris(2-carboxyethyl)phosphine. These initial crystals were improved to diffraction quality by screening many alternative CSN constructs, some of which were identified by mass spectroscopy following limited proteolysis.

Crystals suitable for X-ray studies were from a single CSN variant containing the canonical isoforms of CSN2–CSN8 and CSN1 isoform 2 (as defined by the UniProt database), with truncations to the N-termini of CSN1 (51 residues) and CSN5 (11 residues), and C-terminus of CSN7a

(59 residues). The crystals were rhombohedron-shaped with a diameter of 0.1–0.2 μm. All CSN crystals analysed were of this construct.

Despite searching extensively, reliable conditions for crystal soaking and cryo-protection were not found. Viable cryo-protection was obtained by sweeping each crystal briefly from 15% to 30% (v/v) ethylene glycol in a solution made up to match the crystal-growth solution as closely as possible (artificial mother liquor) and then flash-cooled in liquid nitrogen. Crystal soaking experiments with compounds other than tantalum bromide were carried out in artificial mother liquor. Following heavy-atom treatment, crystals were back soaked for 1–5 min in artificial mother liquor supplemented with 15% (v/v) ethylene glycol before continuing the cryo-protection protocol as for other crystals.

Diffraction data were collected at the Swiss Light Source beamline X06DA with a Pilatus 2M detector (Dectris) or beamline X10SA with a Pilatus 6M detector (Dectris) from crystals cooled to 100 K. Diffraction was typically measured with an intense X-ray beam across single-axis rotations with fine rotational slicing ($\Delta\varphi = 0.05$–0.25°) from multiple positions on the crystal. The diffraction properties varied substantially both from different positions on each crystal and across crystals. Consequently, it was difficult to acquire isomorphous data and elaborate collection strategies combining multiple datasets were rarely useful.

Diffraction images were processed with *XDS*[10] in multiple passes recycling experimental parameters, using optimisation methods described on the *XDSWIKI* (http://strucbio.biologie.uni-konstanz.de/xdswiki). Shadowed regions of the detector were excluded from processing and only the experimental geometry parameters that varied reasonably over the rotation range were refined. Without careful data processing, weak anomalous signal was lost and issues of data-quality were worsened. The value of two statistics from *XDS* were considered as signs of promising anomalous signal: the mean correlation between two random subsets of anomalous differences (CC_{anom}, also known as $CC_{1/2\text{-anom}}$) of $\gtrsim$0.3, and the mean anomalous difference in units of estimated standard deviation (SigAno) $\gtrsim$1.2 for much of the low-resolution range. Crystals containing heavy atoms were susceptible to radiation damage, indicated by processing statistics suggesting substantial anomalous signal but for which substructure searches failed.

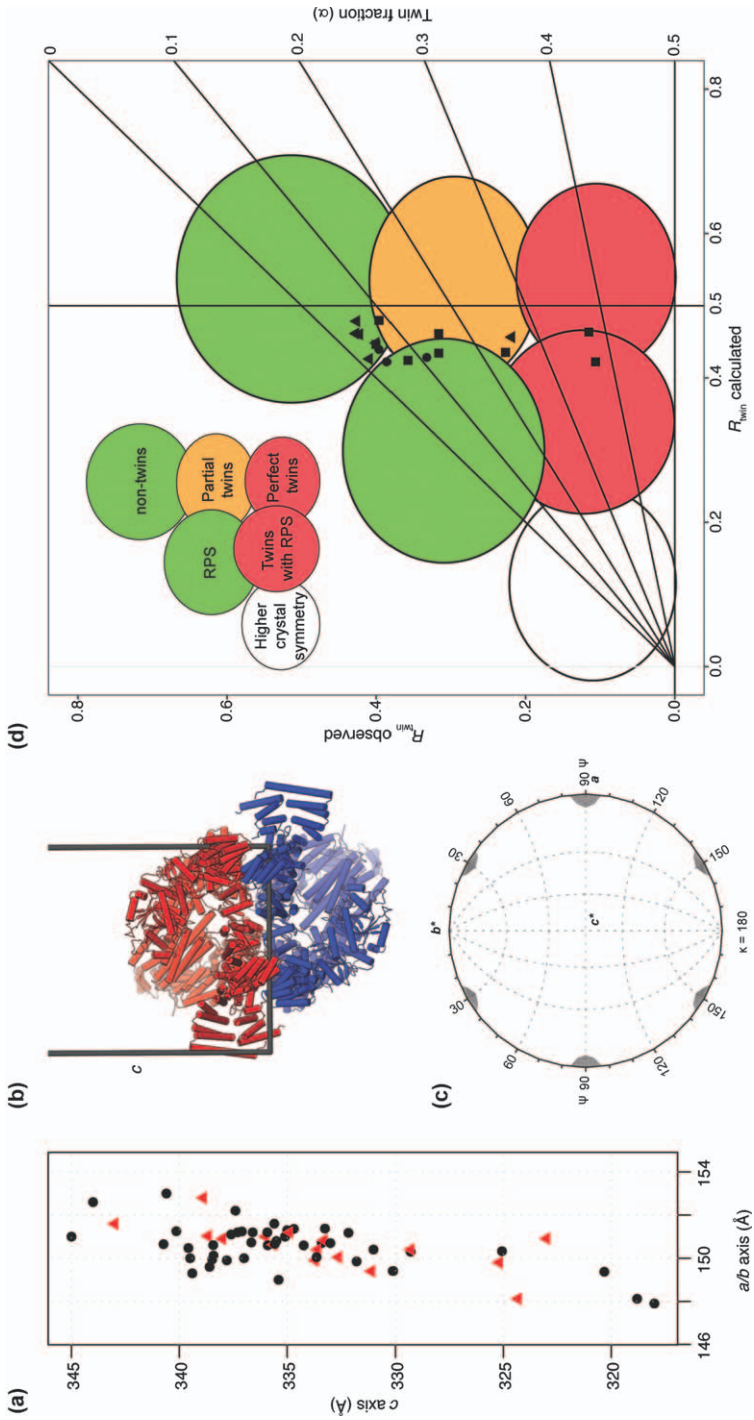
Several scaling protocols using *XDS*/*XSCALE* and *AIMLESS*[11] from the CCP4 suite[12] were applied to each rotational sweep or combination thereof. For the SeMet substituted derivative crystals, the various scaling runs were ranked exclusively by the height and number of anomalous log-likelihood gradient (LLG) map peaks found by MR-SAD with *PHASER*[13] using a common starting model, with the best data being analysed further. The height of the LLG peaks for the CSN5 active site Zn^{2+} ions were a useful reference point for this evaluation. Most often the best results were obtained by supplementing the scaling by the *CORRECT* step of *XDS* with scaling by *AIMLESS* using its secondary beam scaling model without *B*-factor correction. Averaged intensities were converted to structure factor amplitudes with *TRUNCATE*[14] for

phasing and *XDSCONV* for the refinement data sets, applying a flat prior distribution for twinned data. R_{free} sets of reflections for cross-validation were selected with twinning and rotational pseudo-symmetry (RPS) considered. Processed diffraction data were analysed with *POINTLESS*[11] and *phenix.xtriage* from the *PHENIX* suite.[15]

7.3 Crystal Characterisation

X-ray diffraction from CSN crystals was typically poor quality and was restricted to a maximum resolution of only ~6–7 Å. The unit cell parameters varied substantially; the length of the *c* axis by ~27 Å and *a*/*b* axes by ~5 Å (Figure 7.2a). The crystals were clearly trigonal and the pattern of systematic absences consistently indicated a threefold screw axis. Data from most crystals merged best in the enantiomeric space groups $P3_1$ or $P3_2$ and had signs of twinning in the *L*-test.[16] Some data, however, merged convincingly in the higher symmetry space group $P3_121$ or $P3_221$, indicated by an R_{meas} of 0.06 or lower below ~10 Å resolution. Unit-cell content analysis predicted two or three CSNs in the asymmetric unit of the crystal (ASU) in space group $P3_1$ or $P3_2$ (with an approximate solvent content of ~66% or ~50%, respectively) and a single CSN in the ASU in space group $P3_121$ or $P3_221$ (approximately 66% solvent content). Self-rotation function analysis of data in space group $P3_1$ or $P3_2$ revealed a twofold rotational axis perpendicular to the crystallographic threefold screw axis (Figure 7.2c). Before untwinned data were obtained and progress was made in phasing, we were unable to distinguish whether this twofold rotation was (i) a crystallographic symmetry element generating crystals belonging to space group $P3_121$ or $P3_221$ or (ii) a pseudo-crystallographic non-crystallographic symmetry (NCS) rotation (rotational pseudo-symmetry; RPS) or (iii) hemihedral twinning (with twin law $k,h,-l$) or (iv) a combination of these. All crystals were later determined to belong to space group $P3_1$, contain two CSN complexes in the ASU (~66% solvent), and to be affected variably by RPS and twinning (with the exception of two heavy-atom soaked crystals belonging to space group *P*1, which are discussed in Section 7.10) (Figure 7.2d). Twinning and RPS compounded by the limited resolution of the diffraction data made the point group of the crystals ambiguous, complicating the assignment of the correct space group.

RPS and twinning presented different challenges to determining the structure. RPS reduced the power of NCS averaging for phase improvement. RPS correlated reflections related by the twofold NCS axis perpendicular to the threefold crystallographic axis such that the magnitude of a given *h*,*k*,*l* is approximately equal to $k,h,-l$, producing an *R*-factor between these reflections (R_{twin}) of 0.43 at best across all data sets (Figure 7.2d). An R_{twin} of 0.5 is expected for uncorrelated (independent) reflections.[17] Because the NCS mates are not completely independent, there was limited alternative sampling of the molecular transform that could be exploited for phase improvement by electron density averaging.[18] This was especially notable at low resolution where the NCS appeared crystallographic (Figure 7.2c). NCS


(a)
c axis (Å)
a/b axis (Å)
(b)
c
(c)
b*
c*
a*
κ = 180
(d)
R_{twin} observed
R_{twin} calculated
Twin fraction (α)
non-twins
Partial twins
Perfect twins
RPS
Twins with RPS
Higher crystal symmetry

averaging, however, was found to be beneficial in favourable cases. Retrospective analysis of ~4 Å untwinned data with negligible RPS (crystal *c*318; PDB entry 4D18), showed that in a test with prime-and-switch density modification of the electron density from refinement of an early model (that had a correlation coefficient (CC) to the final map of 0.66) in *RESOLVE*,[19] applying NCS averaging with the final optimised groups, as compared to without, increased the CC from 0.71 to 0.81 with respect to the final map.

Twinning introduced more serious problems than RPS to structure solution. Unlike RPS, twinning in CSN crystals interfered with the anomalous signal by correlating Bijvoet pairs. This led to challenges in measuring anomalous data from weak derivative crystals. Tests with 3.8 Å resolution synthetic data calculated for crystal *c*343 (PDB entry 4D10) twinned with a fraction (α) of 0.36 (twin law $k,h,-l$), as in the real data, reduces the mean anomalous difference Fourier peak height for the active site Zn^{2+} ions by 17%. Despite this, peaks for the CSN5 active site Zn^{2+} ions were found for crystal *c*343 at 9.5 and 8.5 times the r.m.s level in an anomalous LLG map calculated in *PHASER*, demonstrating that anomalous data from twinned CSN crystals were useful for substructure identification. Twinning also complicated structure determination by obscuring the electron density for a conformationally variable subunit (CSN4) and contributed to crystal inconsistency (non-isomorphism) by convoluting the intensity of each reflection with a twin that varied across data sets. The crystals typically had an apparent α of 0.35–0.45. Applying a consistent indexing scheme across highly twinned data was also problematic. Crystallization in a trigonal space group with RPS explains why the crystals readily formed hemihedral twins wherein the twin and NCS axes were nearly coincident (Figure 7.2b).[17]

Figure 7.2 Crystal characterisation. (*a*) Scatter plot showing the variation of unit-cell dimensions across 60 CSN diffraction data sets. Crystals with apparent $P3_1$ or $P3_2$ space group symmetry are plotted as black circles, and crystals with apparent $P3_121$ or $P3_221$ space group symmetry are plotted as red triangles. (*b*) The asymmetric unit of a crystal *c*318 (PDB entry 4D18) viewed across the *a*/*b* plane showing the two CSN complexes in cartoon mode related by twofold rotational pseudo-symmetry (RPS). The unit-cell axes are shown as solid grey lines. (*c*) Self-rotation function analysis. The $\kappa = 180^\circ$ section of the native Paterson self-rotation function calculated for the untwinned native data set used for MIRAS phasing (space group $P3_1$) in the resolution range 6.6–20 Å, indicating 32 pseudo-symmetry. The peaks are at 77% the height of the origin peak. (*d*) Twinning and rotational pseudo-symmetry analysis of the data sets contributing to the original structure determination by RvR scatter plot (Lebedev *et al.*[17]) for twin law $k,h,-l$. R_{twin} values from *phenix.xtriage* are given for non-anomalous data and from the final model after rigid body fitting against each crystal. Native crystals are plotted as black triangles, heavy-atom-soaked derivatives as black circles, and the reference data sets for the eight different SeMet-substituted derivatives as black squares.
Panels a–c adapted under a Creative Commons License (https://creativecommons.org/licenses/by/2.0) from Bunker.[2]

Twinning was only modelled in refinement of crystal *c*343 with *REFMAC* and ignored for other procedures and for other data sets.

Finding twinning was suppressed in crystals propagated by seeding allowed untwinned data to be collected. All untwinned crystals were grown by seeding. Repeated generations of streak seeding with a cat's whisker were performed as described previously[20] modified to CSN conditions. In addition to its importance as a method for overcoming twinning, repeated seeding was the only technique found to consistently improve the diffraction quality of the crystals, eventually allowing data to ∼4 Å resolution to be collected. Cross-seeding from native crystals was essential to grow selenomethionine (SeMet)-substituted crystals.

Because phase improvement by NCS averaging was limited, we exploited crystal non-isomorphism for phase improvement by cross-crystal averaging. This proved key to determining the structure. We suspected that crystal variability was strongly influenced by handling but had not found a method to control non-isomorphism. This would have enabled us to systematically induce non-isomorphism to generate good-quality data sets for cross-crystal averaging or more successfully combine isomorphous data from multiple sweeps. Seeking to understand the changes the crystals underwent before data collection at 100 K and to control non-isomorphism, we carried out room temperature diffraction and dehydration experiments using the Humidity Control device (HC1)[21] at the Diamond Light Source, UK. The unit cell dimensions at room temperature were $a = b = \sim 152$ Å, $c = \sim 344$ Å, the largest measured for CSN. Dehydration reduced the crystal volume, replicating the type of unit-cell variation found for cryo-cooled crystals. A few degrees of data were obtained at room temperature. Bragg peaks were visible to ∼8 Å resolution and appeared much sharper than at 100 K, suggesting that flash-cooling was damaging to the diffraction quality of the crystals. Radiation damage at room temperature was severe and we were unable to flash-cool dehydrated crystals to collect sufficient data for analysis.

We instead employed dehydration as a post-crystallization treatment by replacing the well solutions of successful hanging drop vapour diffusion crystallization experiments with various concentrations of PEG 6000. Dehydration screening with 20% PEG 6000 and overnight equilibration yielded one crystal with a *c* axis of 318 Å (crystal *c*318; PDB entry 4D18) that diffracted to 4.1 Å resolution. Additive screening yielded a crystal with a *c* axis of 343 Å (crystal *c*343) that grew spontaneously in a condition supplemented with 10 mM urea. Crystal *c*343 was partially twinned ($\alpha = 0.36$) with RPS and diffracted to 3.8 Å resolution, the highest obtained for CSN and this led to the other finalised structure (PDB entry 4D10). The extreme unit cell variants, *c*318 and *c*344, were complemented by a crystal with a *c* axis of 337 Å (crystal *c*337), which yielded high-multiplicity data to 4 Å resolution. These three non-isomorphous unit-cell variants (crystals *c*318, *c*337 and *c*343), became the focus of refinement and model building.

Though mitigated by optimising the crystals over repeated generations of seeding, problems of crystal instability, non-isomorphism and twinning

were never completely resolved. Some 1400 crystals were analysed to circumvent these issues, from which five native, four heavy atom-soaked derivatives, and a redundant set of eight SeMet-substituted derivatives contributed to the two finalised models.

7.4 CSN4

We anticipated that model building for CSN would be challenging with the low-resolution diffraction data (~4 Å) we had obtained and therefore also set out to determine high-resolution structures of individual subunits to facilitate interpretation of the complex. These attempts produced a structure of human CSN4 on its own, with its C-terminal α-helices deleted and N-terminal Strep(II) affinity tag retained (PDB entry 4D0P). Diffraction data for phasing were collected in-house with a copper rotating anode X-ray generator. Structure determination by SIRAS phasing using a mercury derivative and refinement at 1.6 Å resolution was straightforward. Securing a high-quality model of CSN4 was fortunate because we were unable to interpret the electron density for the N-terminal region of CSN4, which is the most variable part of crystallized CSN, without external structural information.

7.5 Heavy-atom-soaked Derivative Crystals

All initial X-ray diffraction analysis of crystals treated with simple heavy atom salts had been unsuccessful. Considering heavy atom treatment had consistently resulted in crystals with low diffraction quality and the ASU was predicted to contain ~700 kDa of protein, we prioritised preparing cluster-compound derivatives. After a long search, a heavy atom derivative was produced by incubating a crystal in its original drop for one week with a few grains of tantalum bromide. A single wavelength anomalous diffraction (SAD) data set was collected for this crystal to ~9.5 Å resolution that was not isomorphous with other crystals. The processing statistics had promising signs of strong anomalous signal (overall $CC_{1/2\text{-anom}}$ of 0.69). Anomalous Patterson analysis, however, revealed a single tantalum bromide position, an unfortunate situation for SAD phasing. A single heavy atom substructure in a polar space group such as $P3_1$ or $P3_2$ is centrosymmetric. SAD phasing is complicated by centrosymmetric substructure because the noise convoluted with the true electron density map that is inherent to the technique is not random, in contrast to non-centrosymmetric substructures for which this noise is random.[22] This proved fatal to the density modification stage of SAD phasing, which is required to identify the true solution and extract the electron density for the crystal. Although we were unable to advance structure determination with the tantalum bromide derivative data alone, it provided a means to cross-validate the results of initial phasing when a more tractable heavy atom derivative was found.

Carrying on the search, a promising derivative was prepared by soaking crystals with deca-ammonium paratungstate (Jena Bioscience), a distinctly ellipsoidal W_{12} cluster compound ($[H_2W_{12}O_{42}]^{10-}$; W_{12}). Sites were found with *phenix.hyss*[15] using anomalous data to 9.2 Å resolution in a search spanning a range of low-resolution cut-offs. Discovering the number of W_{12} cluster sites could be manipulated by changes to the duration and concentration of the compound used for crystal soaking, which had been an ineffective approach with tantalum bromide, two derivatives suitable for phasing were prepared (designated W_{12}-I and W_{12}-II). Two well-occupied W_{12} sites were found in a crystal soaked overnight with 0.25 mM W_{12} using data collected at the tungsten L_{III} absorption peak (W_{12}-I; $f' = -16$, $f'' = 24$) (Figure 7.3a). A total of four W_{12} sites were found for a crystal soaked overnight with 1 mM W_{12} using data collected at the inflection point of the tungsten L_{III} absorption edge (W_{12}-II; $f' = -28$, $f'' = 15$); two equivalents of W_{12}-I and two unique at ~20% of the occupancy of the first pair (Figure 7.3a). Both W_{12} derivative crystals diffracted to ~6.6 Å resolution with a 2 $\langle I/\sigma I \rangle_{mrgd}$ cut-off. W_{12}-I had a $CC_{1/2\text{-anom}}$ of 0.85 and SigAno of 2.46 and W_{12}-II a $CC_{1/2\text{-anom}}$ of 0.84 and SigAno of 2.50 from the low-resolution limit to 9.2 Å, beyond which $CC_{1/2\text{-anom}}$ dropped sharply below 0.3.

The two major W_{12} positions in each derivative were NCS equivalents wedged between the N-terminal arm of CSN2 and CSN8 in a neighbouring ASU (Figure 7.3b). The two weaker sites in W_{12}-II were also NCS equivalents nestled beside CSN3, though neither NCS relationship was recognised until the election density improved after initial phasing (Figure 7.3a).

Substructures were unable to be found for many data sets collected from heavy atom-soaked crystals. These were revisited as structure determination progressed using MR-SAD with *PHASER,* and only a single data set yielded a heavy atom substructure; a derivative prepared with a *p*-chloromercurial benzoic acid. This mercury derivative, which diffracted to ~7.5 Å resolution, was partially twinned ($\alpha = 0.24$) and did not have promising signs of anomalous signal in its processing statistics, confirmed the location of 21 cysteine residues in the final model (Figure 7.4).

7.6 Initial Phasing

Combinations of native and derivative data sets were evaluated for their potential for phasing, guided by the analysis steps of *autoSHARP*[23] and criteria previously outlined.[24]

The two W_{12}-cluster derivatives were combined with an isomorphous, untwinned native data set for MIRAS phasing. After relative scaling with *SCALEIT* though *autoSHARP*, W_{12}-I and W_{12}-II had unweighted cross-crystal *R*-factors on amplitudes to the native data set of 0.183 and 0.173, respectively, and an *R*-factor of 0.135 to each other across the full resolution range (50–6.6 Å). The weighted *R*-factor to the native data set was larger than the *R*-factor across the resolution range 50–7.7 Å for both derivatives. Phasing

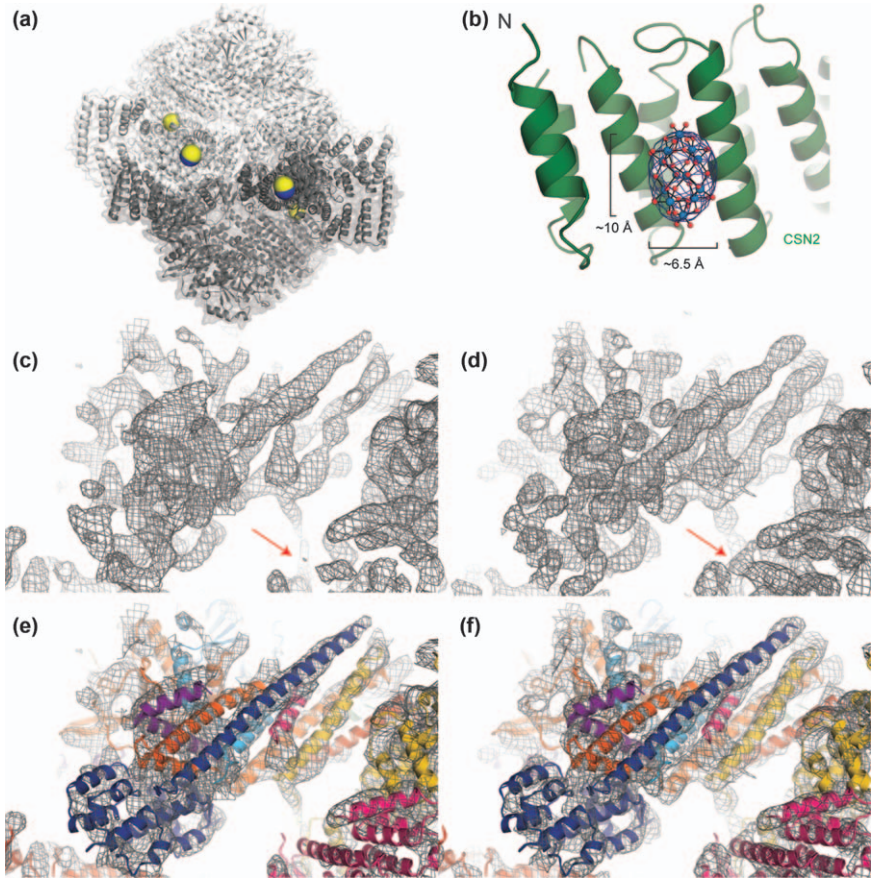


Figure 7.3 Experimental phasing. (*a*) A view of an asymmetric unit of the crystal showing the NCS copies of CSN in cartoon mode in contrasting shades of grey and the paratungstate cluster (W_{12}) positions as spheres for the two correlated derivatives used for MIRAS phasing coloured blue for W_{12}-I and coloured yellow for W_{12}-II. (*b*) A close-up view showing for the major W_{12} position beside CSN2, which was found for both derivatives with an NCS equivalent. An anomalous Difference Fourier map calculated for W_{12}-I after rigid body refinement of the final CSN model with *phenix.refine* at 6.8 Å resolution is shown as blue mesh at a contour level of 10 r.m.s.d (peak height 19.5 r.m.s.d), emphasising the ellipsoidal shape of the LLG density. A single paratungstate ion is fitted into the peak as a rigid body and drawn in ball and stick mode. The length of the major axis of the density is indicated. (*c*) A segment of the MIRAS-phased experimental electron density map surrounding CSN7 calculated at 9 Å resolution and contoured at 1 r.m.s.; (*d*) the same segment after density modification by solvent flipping with phase extension to 6.6 Å resolution. The red arrows highlight a region improved by density modification. (*e*) As in (*c*), and (*d*) as in (*f*), the same view with the final CSN model embedded in the electron density map.
Panels *a* and *c*–f adapted under a Creative Commons License (https://creativecommons.org/licenses/by/2.0) from Bunker.[2]

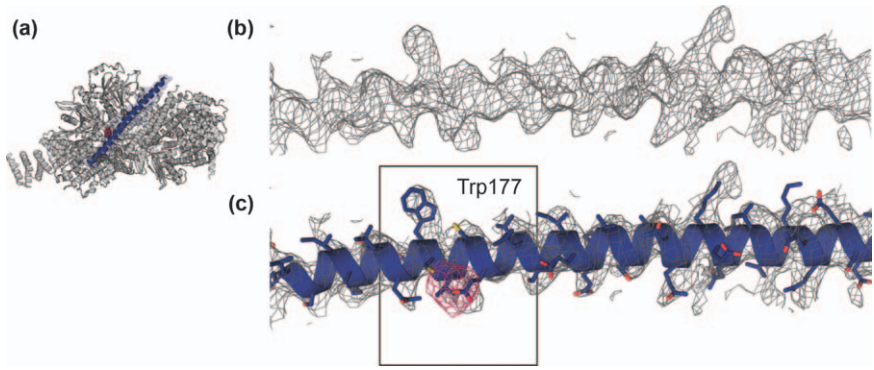


Figure 7.4 Sequence assignment and validation of the 50-residue (70 Å long) CSN7 C-terminal α-helix. (*a*) The location of the α-helix (highlighted in blue) within CSN (white in cartoon mode). Its initial sequence assignment was anchored on a prominent side chain bump for Trp177 in a 4 Å resolution electron density map calculated by cross-crystal averaging across the three refinement data sets for crystal *c*343 with *DMMULTI*. (*b*) A segment of this electron density map (grey mesh) contoured at 1 r.m.s. with a radius of 4 Å around the displayed model. (*c*) The same view as (*b*) with the final model embedded (shown in cartoon mode with side chains in blue). The sequence register was confirmed with a peak found for a modified cysteine residue neighbouring Trp177 in an anomalous LLG map (pink mesh; peak height of 6.3 r.m.s.) calculated for a mercury derivative crystal. Moreover, Trp177 had been shown biochemically to be crucial for the interaction of CSN7 with CSN6.[47] The side chain of CSN7 Trp177 is buried in a hydrophobic pocket created by CSN6 in CSN, explaining the biochemical findings.
Reprinted under a Creative Commons License (https://creativecommons.org/licenses/by/2.0) from Bunker.[2]

was initiated with *SHARP*[25] with one of the two sites found for W_{12}-I, using LLG map peaks found for its second site to confirm the veracity of both. The heavy atom model was completed for both derivatives by iterative interpretation of the LLG maps first as single atoms then as spherically averaged descriptions of W_{12} parameterised through the SPHCLUSTER keyword in *SHARP*. Only the position and occupancy of the W_{12} clusters were refined, with their *B*-factors fixed at 150 $Å^2$ or 300 $Å^2$. The scaling and *B*-factor-like parameters were poorly determined in *SHARP*, particularly before the W_{12} model was complete, giving rise to inflated estimates of phase quality. Fixed values of the global non-isomorphism parameters (NISO_BGLO and NANO_BGLO in *SHARP*) were therefore applied in cycles with an incomplete heavy atom model. By designating the derivative W_{12}-II as the reference for MIRAS phasing, we avoided correlated non-isomorphism in *SHARP*, which can also lead to over-estimated phase quality.

Phases calculated from spherically averaged cluster positions were restricted to the approximate diameter of the W_{12} sites (~9.2 Å resolution), substantially lower resolution than the MIRAS phasing data sets

(~6.6 Å resolution). The LLG map peaks for the major sites were ellipsoidal (Figure 7.3b), suggesting the cluster had a preferred orientation for binding that could potentially be exploited for phasing. Several approaches aimed at extending the phased resolution beyond ~9 Å by modelling the fine structure the W_{12} sites for phasing were attempted with limited success. These included placing multiple sites closely together in the LLG map peaks, reducing the radius of the cluster for spherical averaging as well as treating the tungsten atoms separately and phasing from the W_{12} framework of the cluster fitted as a rigid body or from many tungsten atoms placed individually. The best phases, with an overall figure-of-merit (FOM) of 0.78, were obtained from a single spherically averaged W_{12} placed at each site with a slightly smaller radius (4.5 Å) than expected from its crystal structure (~4.75 Å).

After density modification by solvent flipping with *SOLOMON*[26] operated through its interface in *SHARP*, the correct hand of the W_{12} substructure was readily distinguished by the presence of bundles of tubular density, indicative of α-helices, in space group $P3_1$ and fragmented electron density for space group $P3_2$ (Figure 7.3). The key parameters to optimise in *SOLOMON* were the low-density truncation level for the protein region and solvent content.

Although little phase improvement was derived from NCS averaging, it was nevertheless developed and carried out in the hope it would be eventually prove useful. The twofold NCS operator was identified in the *SOLOMON* map with *phenix.find_ncs_from_density*[27] and applied for averaging with *RESOLVE.* The combination of density modification with *SOLOMON* and averaging with *RESOLVE* generated phases to 6.6 Å resolution.

As the primary source of structural information for CSN, care was taken to produce the best possible initial electron density map to support subsequent structure determination steps. Progress and parameter selection throughout initial phasing was guided by cross-validation against the peak height in anomalous difference Fourier maps calculated for the tantalum bromide derivative. This provided an indication of the phase error independent of the FOM and other internal estimates of phase quality. Although FOM is theoretically the mean cosine of the phase error, in practice the FOM calculated from phase probabilities generally underestimates their error and can be an unreliable statistic for ranking phase quality.[28]

7.7 Subunit Identification and Selenomethionine Phasing

Because of the similarity within the two subunit families (the six PCI subunits and two MPN subunits), the splayed structure of CSN (Figure 2) and the nature of its crystal packing interactions, the two biological complexes in the ASU were unable to be defined in the low-resolution electron density maps from initial phasing.

We produced CSN recombinantly by infecting insect cells with separate viruses for each subunit. To identify the individual subunits and map the layout of the complex, we exploited the remarkable flexibility of the CSN assembly pathway, which allowed purification of the complete complex from complementary sub-assemblies mixed at lysis. Several selectively substituted SeMet derivatives were prepared, and the selenium sites found for these crystals were used to define its subunit arrangement and extent in the initial electron density maps. We also aimed to derive phase information from the SeMet derivatives to improve the electron density and, therefore, performed combinatorial labelling of subunits in an attempt to give rise to adequate signal for phasing. Diffraction data were obtained for four derivatives with single SeMet protein substitutions (CSN1; CSN2; CSN3; and CSN8) and four combinatorial derivatives (CSN1 and CSN4; CSN2 and CSN4; CSN2, CSN3, CSN5, CSN8; and CSN2, CSN3, CSN5, CSN6, CSN7, CSN8).

A sphere containing the approximate volume of an ASU was excised from the electron density map from initial phasing and prepared as a molecular replacement (MR) search model using *phenix.cut_out_density*. The electron density was distributed to the SeMet derivative crystals by rigid body refinement (or MR) with *PHASER*. Protocols describing the use of density maps as MR search models with *PHASER* are given by Jackson and coworkers.[29] The positioned electron density was used to initiate an anomalous substructure search using MR-SAD in *PHASER*. The electron density was improved with cross-crystal averaging with a 6 Å resolution cut-off, combining the current best Se-SAD phases and MIRAS phases with two non-isomorphous native data sets in *phenix.multi_crystal_average.* The procedure was then repeated starting with the improved electron density map.

LLG map peaks with a root mean square (r.m.s.d.) above the mean value of 6 were assigned as heavy atoms automatically by *PHASER*. Importantly, the LLG maps after automatic interpretation were evaluated manually, and peaks above ~4 r.m.s.d. were assigned as heavy atoms if they were found in multiple crystals. Two Zn^{2+} ions, one for each CSN5, were consistently found for derivative crystals, and were confirmed with native data. This procedure plateaued after three iterations, having found 82 of the 176 selenium atoms expected in the ASU.

Most selenium positions were found in derivatives in which all subunits except CSN1 and CSN4 were labelled with selenium, largely from one exceptional crystal that diffracted to 4.8 Å resolution and gave phases to ~8.0 Å resolution assuming a 0.3 FOM cut-off. The other derivatives were of considerably lower quality but crucial for identifying the subunits. Based on the ensemble of selenium and zinc sites each subunit was identified and the biologically relevant ASU defined.

7.8 Initial Model Building

An initial poly-alanine model was built interactively with *COOT*[30] into the best electron density map from subunit identification and SeMet phasing

(at 6 Å resolution). Fragments of the models for human CSN4 and CSN5, structures of CSN1, CSN6 and CSN7, and homology models of CSN2 and CSN8 were positioned, manually guided by the constellation of selenium sites as markers for methionine residues and secondary structure predictions from *PSIPRED*.[31] Homology models were constructed with *PHYRE2*[32] and *I-TASSER*.[33] Ideal poly-alanine α-helices were docked into tubular electron density not described by model fragments.

At this point, many refinement and density modification approaches were trialled in an attempt to improve the interpretability of the electron density. The successful approach that emerged was to repeatedly extend and rebuild the unrefined model into electron density calculated for a 4 Å resolution data set (crystal *c*337) by refinement with *autoBUSTER* and subsequent prime-and-switch density modification with NCS averaging in *RESOLVE*. At first, the major fragments were optimised as rigid bodies. Missing atom modelling was enabled using the cross-crystal averaged phases from subunit identification in refinement and to define the distribution of unmodelled atoms (as a non-uniform prior). The cross-crystal averaged phases were later abandoned and it became useful to include positional refinement with local structure similarity restraints (LSSR) between NCS mates and to the high-resolution models of human CSN4 (PDB entry 4D0P) and CSN5 (PDB entry 4F7O).

Model building was carried out in *COOT* with tight geometry restraints supplemented by a Ramachandran plot and secondary structure restraints as appropriate. Progress in initial model building stalled with a fragmented model containing approximately half the number of atoms of the final model, prompting us to search for different approach to complete the model.

7.9 Model Completion

A poly-serine version of the model from initial building of crystal *c*337 was transferred to two non-isomorphous native crystals (*c*318 and *c*343) by rigid body refinement using *autoBUSTER* with one group per subunit. The electron density produced by rigid body refinement and various ensuing refinement attempts was, however, uninformative. This was attributable to large-scale conformational changes within CSN accompanying changes to the unit-cell volume, that were unable to be accommodated by conventional refinement, as well as the inaccuracy of the current model. Progress was, however, made by following rigid body fitting with deformable elastic network refinement (DEN) with *CN*S.[34] DEN produced good-quality electron density maps that clearly revealed conformational differences in CSN among the structures. DEN was performed using the input model as reference with tight NCS restraints and parameter optimisation as described by Schröder *et al.*[35] The DEN results showed that the conformation of CSN4 varied dramatically across crystals and between NCS mates and from which a more effective rigid body decomposition could be defined for refinement and

electron density averaging. Rigid domains were identified manually in *COOT* and with *DYNDOM3D*.[36] The secondary structure of the DEN models was severely disrupted and we therefore replaced them with earlier models before continuing.

The model was then extended and refined iteratively across the three non-isomorphous unit-cell variants (*c*318, *c*337 and *c*343). The untwinned structures, *c*318 and *c*337, were refined using *autoBUSTER*, with NCS and reference (target) model restraints applied as LSSRs.[37] The twinned structure, *c*343, was refined using *REFMAC*[38] against structure factor amplitudes, with local NCS restraints and reference model restraints generated by *PROSMART*.[39] Co-refinement was carried out using reference model restraints generated across crystals to share structural information as much as possible among the three unit-cell variants. For *autoBUSTER*, restraints were typically generated across crystals *c*318 and *c*337, and *c*337 and *c*343, to regions of the model not corresponding to the high-resolution CSN4 and CSN5 models. *REFMAC*, however, accepted restraints generated across multiple structures. This procedure allowed the three structures to be readily compared and facilitated cross-crystal averaging. Co-refinement also stabilised refinement by improving the effective observation-to-parameter ratio, as judged by a reduction of the $R_{\mathrm{free}} - R_{\mathrm{work}}$ gap, and limited divergence of the models not warranted by the diffraction data. Variation of the crystal packing interactions across crystals improved the interpretability of regions of the electron density in each unit-cell variant with respect to the others, notably for the periphery of the complex. By facilitating comparison with untwinned structures, co-refinement also provided an internal control for model bias in twin refinement of *c*343, which we approached cautiously.

Maximum likelihood twin refinement of *c*343 with *REFMAC* gave the most interpretable electron density for model building. Modelling the twin in refinement reduced $R_{\mathrm{work}}/R_{\mathrm{free}}$ dramatically (by ~7%) but provided only a subtle improvement to the quality of the electron density map, as expected for a structure wherein the NCS and twinning axes are nearly coincident (Garib Murshudov, personal communication). Crucially for model building, the clarity of the electron density for some side chains was enhanced by twin refinement.

NCS averaged prime-and-switch density modified versions of the electron density from refinement were generated with *RESOLVE,* applying the rigid bodies identified earlier. These maps were then averaged across crystals with NCS averaging *phenix.multi_crystal_average* and *DMMULTI* from the CCP4 suite. The resulting electron density maps were used for model building in combination with the electron density directly from refinement. The final averaging procedure was carried out across the three non-isomorphous refinement data sets and the high-resolution CSN4 structure with CSN divided into 12 groups (Figure 7.5). Negative *B*-factor sharpening of the structure factor amplitudes before density modification and map coefficients from refinement were broadly applied to enhance the electron density for side chains.

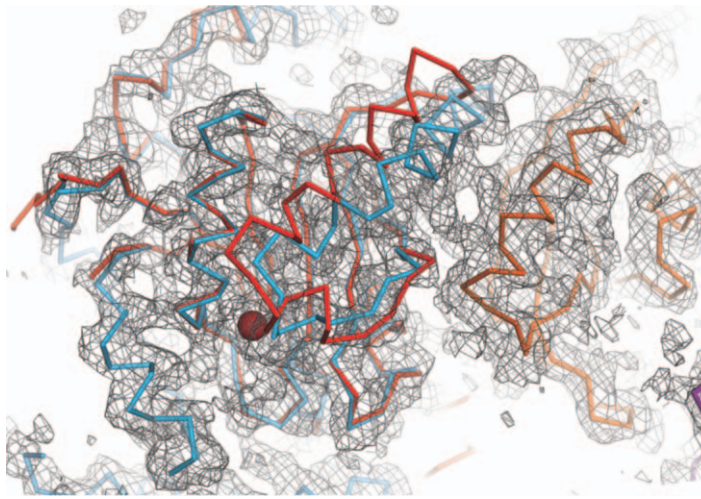


Figure 7.5 Representative electron density for CSN. A portion of the electron density map (grey mesh) surrounding the active site of CSN5 (cyan) calculated at 4 Å resolution and contoured at 1 r.m.s. the density for crystal *c*337 by cross-crystal averaging with NCS, averaging across the three non-isomorphous unit-cell variant refinement data sets with *phenix.multi_crystal_average*. The structure factor amplitudes were sharpened by -134 Å^2 to an overall B-factor of 40 Å^2. Cα traces are shown for CSN5 (cyan) in superposition with the high-resolution structure of CSN5 (red) (PDB entry 4F7O), CSN6 (orange). CSN5 assumes an auto-inhibited conformation in CSN, which was not found for the crystal of the isolated CSN5 fragment. This results in conformational differences at the active site (indicated by a red sphere for the active site Zn^{2+} ion) of the two CSN5 models.
Reprinted under a Creative Commons License (https://creativecommons.org/licenses/by/2.0) from Bunker.[2]

The refinement protocol that we used during the model completion was conservative. The weight of the X-ray to geometry terms was determined automatically in *REFMAC5*, and a fixed value used for *autoBUSTER* that maximised electron density map quality, which we assessed using several omit regions that were withheld from model building.

Owing to our growing interest in the conformational dynamics of CSN discussed below, we finalised the models for the extreme unit-cell variants, crystals *c*318 and *c*343.

Near completion of *c*343, a local installation of *PDB_REDO*[40] was modified to incorporate external structural restraints and suitable TLS, jelly-body and twin refinement parameters. This was used to determine optimised weights for positional and *B*-factor refinement, and solvent masking parameters. A high-resolution limit of 3.8 Å for *c*343 (2 $\langle I/\sigma I \rangle_{\mathrm{mrgd}}$ at 4.1 Å resolution) was decided by paired-refinement[41] with *PDB_REDO*.

Similarly, in the final cycles of refinement for *c*318 with *autoBUSTER* an optimal X-ray weight, *B*-factor refinement scheme and through-bond *B*-factor correlation weight were selected from search values in refinement by

assessing the results by R_{work}/R_{free}, LLG_{free}, and geometry validation with *MOLPROBITY*,[42] analogous to the *PDB_REDO* procedure. A high-resolution limit of 4.08 Å for *c*318 (2 $\langle I/\sigma I \rangle_{mrgd}$ at 4.5 Å resolution) was decided by paired-refinement with *autoBUSTER,* using R_{work}/R_{free} calculated by *REFMAC* in a similar manner to *PDB_REDO*.

In the final stages of model building, both *c*318 and *c*343 were refined with *phenix.rosetta_refine*.[43] Based on these refined models, side chain rotamers were selected for residues poorly defined by electron density and lacking a counterpart in a high-resolution reference structure. This allowed information from a distinct refinement method (molecular dynamics) to be considered in model building. All non-hydrogen atoms were included for the modelled regions of both finalised structures.

Supplemental restraints were included in refinement for the coordination geometry of the CSN5 active site Zn^{2+} ion and secondary structure, including intermolecular hydrogen bonds across the β-sheet of the PCI ring. Secondary structure restraints were applied as hydrogen bond-type distance restraints automatically generated by *phenix.secondary_structure_restraints* and were edited and expanded manually in *COOT*. These were applied directly in *REFMAC* combined with reference model restraints from *PROSMART,* and for *autoBUSTER* converted to half-harmonic restraints activated beyond a threshold distance of 3 Å. TLS parameters were included for refinement of both finalised structures based on the domains used for electron density averaging, with groups for the CSN5–CSN6 heterodimer and the α-helical bundle spanning multiple subunits.

The data processing and refinement statistics for *c*318 and *c*343, and human CSN4, and data processing statistics for *c*337, and the MIRAS phasing data sets, are published elsewhere.[1]

Having characterised striking conformational differences in CSN in the course of determining its crystal structure, we became interested in their functional significance. This developed from observations that: (i) an autoinhibited configuration of CSN was captured in the crystals that must be reconfigured for catalysis (Figure 7.5), (ii) preliminary structural information for CSN-substrate (CRL) complexes suggested that for catalysis to occur, CSN must also undergo large-scale conformational change,[1,44] and (iii) there were biochemical data showing that these conformational changes were important for substrate specificity.[1] The role of the conformational changes in the function of CSN became a focus of a follow-up structure-function study of its complexes with various CRL substrates,[3] leading us to extend the analysis of CSN conformers.

7.10 Analysis of CSN Conformational Dynamics Aided by a *P*1 Crystal Form

The unusual properties of two data sets collected from crystals soaked in gold salts were recognised in a review of the crystallography after

completion of the CSN crystal structure. There were no indications these crystals were derivatives, and their overall quality appeared low. The diffraction data could only be processed in space group *P*1. The unit cell parameters of $a = 150.6$ Å, $b = 151.0$ Å, $c = 336.8$ Å, $\alpha = 92.3^\circ$, $\beta = 92.4^\circ$, $\gamma = 119.9^\circ$ were clearly related to the usual trigonal crystal form ($a = b = \sim 150$ Å, $c = \sim 336$ Å, $\alpha = \beta = 90^\circ$, $\gamma = 120^\circ$). We hypothesised that in breaking the rotational symmetry found in the trigonal crystals, CSN may have undergone larger-scale conformation changes than previously characterised. This was confirmed when we determined the structure of the better of the two *P*1 crystals. Low-multiplicity data were collected for this crystal, extending to ~6.6 Å resolution with a 2 $\langle I/\sigma \rangle I_{\mathrm{mrgd}}$ cut-off. Weak data extending to 5.5 Å resolution were included for refinement. The structure was determined by molecular replacement with *PHASER* by positioning six copies of CSN (crystal *c*343; chains A–H from PDB entry 4d10) in the unit cell. Segmented rigid-body refinement was carried out with *phenix.refine* using groups established previously, generating a model with $R_{\mathrm{work}}/R_{\mathrm{free}}$ (%) 29.90/32.19. The model was then refined by DEN in *CNS,* incorporating positional and grouped *B*-factor refinement (two per residue) with tight sixfold NCS restraints over the same groups as rigid body refinement, excluding the flexible N-terminal arm of CSN4. A DEN parameter search was conducted with a slow-cool simulated annealing scheme from 3000 K over eight refinement cycles with DEN restraints calculated from the MR solution and maintained throughout. The best DEN model had an $R_{\mathrm{work}}/R_{\mathrm{free}}$ (%) of 25.17/28.69. An important loop linking the CSN6 MPN domain to its bundle α-helices that was not included in the MR search model, but which was clearly defined in NCS averaged electron density calculated in *COOT*, was added to the model from an earlier structure in which it was defined by electron density (see Extended Data Figure 6h in ref. 1). The *B*-factors were set to 75 Å^2 and a TLS-only atomic displacement model across the same groups as before was refined for 10 cycles in *REFMAC* immediately followed by 20 cycles of positional refinement with local NCS restraints in which coordination geometry of the active site Zn^{2+} ions was corrected. The final model, with an $R_{\mathrm{work}}/R_{\mathrm{free}}$ (%) of 25.41/28.21 was deposited in the PDB as entry 4WSN and is described in Cavadini *et al.*[3]

Comparisons across the 10 crystallographic conformers of CSN described in the two trigonal crystals, *c*318 and *c*343 (PDB entries 4D18 and 4D10), and the *P*1 structure (PDB entry 4WSN), revealed the type of conformational change CSN can undergo (Figure 7.6). Knowledge of these movements in CSN was exploited in the interpretation of various cryo-electron microscopy structures of CSN-CRL complexes[3] determined at lower resolution (ranging from 6.4 Å to 8.8 Å) compared with the original crystal structures of CSN, and provided insight into substrate discrimination and catalytic activation of CSN.

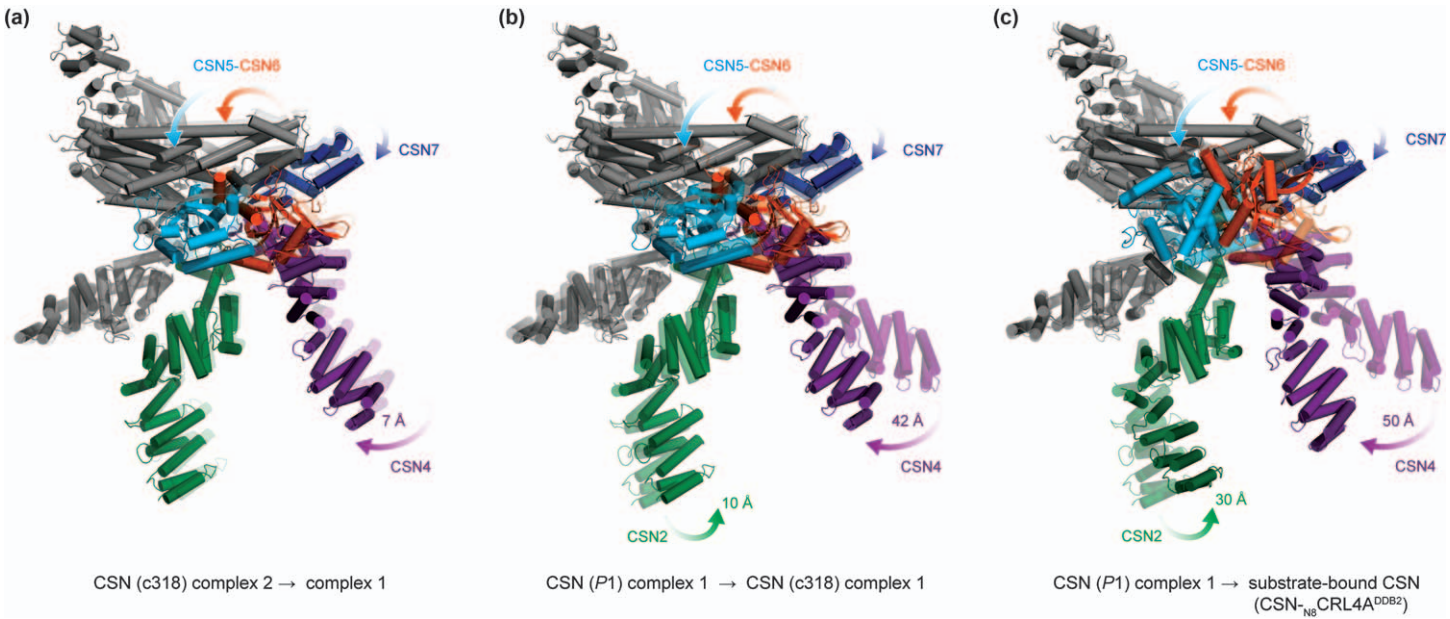


Figure 7.6 Pairwise comparison of CSN conformers. (*a*) NCS mates in crystal *c*318 (PDB entry 4D18). (*b*) The most dissimilar conformers determined by crystallography (*P*1 conformer from PDB entry 4WSN). (*c*) With and without bound substrate, highlighting the dramatic movement of the CSN5–CSN6 dimer (substrate-bound CSN conformer based on EMDB-3316). Comparisons based on superposition of the CSN8 PCI domain. Important domain movements in CSN4, CSN2, CSN5–CSN6, and CSN7 are indicated by arrows. The magnitude of the displacements of the N-termini of CSN4 and CSN2 are indicated in Å if larger than 5 Å.

7.11 Conclusions

Progress in structure determination was gradual, with regular sanity checks required to convince us we were advancing. Structure determination began to look feasible after three non-isomorphous native data sets were collected whose quality stood out from the others, diffracting to interpretable resolution (~4 Å). Initial phasing and substructure searches were re-initiated several times as progress was made using an electron density map or atomic coordinates as the starting model. Model building and refinement was conservative, repeatedly returning to unrefined structures until a successful strategy to complete the models emerged. The final structures, however, rewarded our efforts by providing important insights into the regulation and activity of CSN and the architecture of PCI complexes.

The structure determination unfolded in a manner that reflected the lack of reliable structural information for CSN and PCI complexes at the time. Initial phasing was carried out with only high-quality structures for CSN7 (*A. thaliana* CSN7), a paralogue CSN8 from eIF3 (human eIF3K), explaining our initial focus on experimental phasing and procedures involving electron density models rather than coordinate models.

The *P*1 crystals were discovered after the original crystallographic analysis. The conformational differences across the six NCS copies were found in retrospect to be too complex to model for phase improvement by electron density averaging during initial phasing (Figure 7.6), and the overall resolution too poor for the *P*1 form to be useful for phase improvement subsequently when a coordinate model was available to define the conformational differences. Had methods to refine the shape of electron density models independent of atomic coordinates been available, modelling the conformational changes in CSN across all crystals may have been easier and the power of the cross-crystal averaging procedures potentially increased.

Many difficulties in the original crystallographic analysis can be traced to dramatic conformational variation in CSN4 (Figure 7.6), which broke the symmetry that would have made the NCS crystallographic. Had the NCS been crystallographic, the crystals belonging to space group $P3_121$ would not be able to twin in the same manner. Crystal contacts stabilising CSN4 may have increased the resolution of diffraction, increased isomorphism, and allowed derivatives to be found more easily. Avoiding these crystallographic issues by pursuing a structure of CSN with an N-terminally truncated version of CSN4 would have potentially limited our functional insight, given its roles in substrate recognition and induced-fit activation exposed by the structural analysis.

The co-refinement approach used here was inspired by ideas from the program *MAIN*.[45] Proposals to extend this to simultaneous cross-crystal refinement with electron density averaging[38,46] would have been useful in the case of CSN and likely other challenging structure determinations where it may be desirable to combine information from a variety of sources.

Acknowledgements

The crystal structure of the COP9 signalosome was determined in collaboration with Lingaraju G. Manjappa in the laboratory of Nicolas H. Thomä, both of whom I gratefully acknowledge. I thank Airlie McCoy for many helpful discussions; and Clemens Vonrhein and the BUSTER development group, Randy Read, Tom Terwilliger, Rob Nicholls, Richard Kingston, and Kay Diederichs for advice. The staff of the Swiss Light Source and the Diamond Light Source that supported this work are appreciatively acknowledged. This project was supported by the Novartis Research Foundation, the Swiss National Science Foundation (31003A_144020) and the European Research Council (MoBa-CS #260481).

References

1. G. M. Lingaraju, R. D. Bunker, S. Cavadini, D. Hess, U. Hassiepen, M. Renatus, E. S. Fischer and N. H. Thomä, *Nature*, 2014, **512**, 161–165.
2. R. D. Bunker, *Acta Crystallogr., Sect. D*, 2016, **72**, 326–335.
3. S. Cavadini, E. S. Fischer, R. D. Bunker, A. Potenza, G. M. Lingaraju, K. N. Goldie, W. I. Mohamed, M. Faty, G. Petzold, R. E. J. Beckwith, R. B. Tichkule, U. Hassiepen, W. Abdulrahman, R. S. Pantelic, S. Matsumoto, K. Sugasawa, H. Stahlberg and N. H. Thomä, *Nature*, 2016, **531**, 598–603.
4. J.-H. Lee, L. Yi, J. Li, K. Schweitzer, M. Borgmann, M. Naumann and H. Wu, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 11845–11850.
5. A. Echalier, Y. Pan, M. Birol, N. Tavernier, L. Pintard, F. Hoh, C. Ebel, N. Galophe, F. X. Claret and C. Dumas, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 1273–1278.
6. H. Zhang, Z.-Q. Gao, W.-J. Wang, G.-F. Liu, E. V. Shtykova, J.-H. Xu, L.-F. Li, X.-D. Su and Y.-H. Dong, *FEBS Lett.*, 2012, **586**, 1147–1153.
7. M. Dessau, Y. Halimi, T. Erez, O. Chomsky-Hecht, D. A. Chamovitz and J. A. Hirsch, *Plant Cell*, 2008, **20**, 2815–2834.
8. G. R. Pathare, I. Nagy, S. Bohn, P. Unverdorben, A. Hubert, R. Körner, S. Nickell, K. Lasker, A. Sali, T. Tamura, T. Nishioka, F. Förster, W. Baumeister and A. Bracher, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 149–154.
9. J. Boehringer, C. Riedinger, K. Paraskevopoulos, E. O. D. Johnson, E. D. Lowe, C. Khoudian, D. Smith, M. E. M. Noble, C. Gordon and J. A. Endicott, *Biochem. J.*, 2012, **448**, 55–65.
10. W. Kabsch, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2010, **66**, 125–132.
11. P. R. Evans and G. N. Murshudov, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2013, **69**, 1204–1214.
12. M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. Mccoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton,

H. R. Powell, R. J. Read, A. Vagin and K. S. Wilson, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2011, **67**, 235–242.
13. A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni and R. J. Read, *J. Appl. Crystallogr.*, 2007, **40**, 658–674.
14. S. French and K. Wilson, *Acta Crystallogr., Sect. A*, 1978, **34**, 517–525.
15. P. Adams, P. Afonine, G. Bunkóczi, V. Chen, I. Davis, N. Echols, J. Headd, L. Hung, G. Kapral and R. Grosse-Kunstleve, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2010, **66**, 213–221.
16. J. E. Padilla and T. O. Yeates, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2003, **59**, 1124–1130.
17. A. A. Lebedev, A. A. Vagin and G. N. Murshudov, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2005, **62**, 83–95.
18. G. J. Kleywegt and R. J. Read, *Structure*, 1997, **5**, 1557–1570.
19. T. C. Terwilliger, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2000, **56**, 965–972.
20. R. D. Bunker, J. M. J. Dickson, T. T. Caradoc-Davies, K. M. Loomes and E. N. Baker, *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.*, 2012, **68**, 1259–1262.
21. J. Sanchez-Weatherby, M. W. Bowler, J. Huet, A. Gobbo, F. Felisaz, B. Lavault, R. Moya, J. Kadlec, R. B. G. Ravelli and F. Cipriani, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2009, **65**, 1237–1246.
22. A. J. McCoy and R. J. Read, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2010, **66**, 458–469.
23. C. Vonrhein, E. Blanc, P. Roversi and G. Bricogne, *Methods Mol. Biol.*, 2005, **364**, 215–230.
24. G. Rudenko, L. Henry, C. Vonrhein, G. Bricogne and J. Deisenhofer, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2003, **59**, 1978–1986.
25. G. Bricogne, C. Vonrhein, C. Flensburg, M. Schiltz and W. Paciorek, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2003, **59**, 2023–2030.
26. J. P. Abrahams and A. G. Leslie, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 1996, **52**, 30–42.
27. T. C. Terwilliger, *J. Struct. Funct. Genomics*, 2013, **14**, 91–95.
28. N. S. Pannu, A. J. McCoy and R. J. Read, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2003, **59**, 1801–1808.
29. R. N. Jackson, A. J. McCoy, T. C. Terwilliger, R. J. Read and B. Wiedenheft, *Nat. Protoc.*, 2015, **10**, 1275–1284.
30. P. Emsley, B. Lohkamp, W. G. Scott and K. Cowtan, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2010, **66**, 486–501.
31. D. W. A. Buchan, F. Minneci, T. C. O. Nugent, K. Bryson and D. T. Jones, *Nucleic Acids Res.*, 2013, **41**, W349–W357.
32. L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass and M. J. E. Sternberg, *Nat. Protoc.*, 2015, **10**, 845–858.
33. Y. Zhang, *BMC Bioinform.*, 2007, **9**, 40.
34. G. F. Schröder, M. Levitt and A. T. Brunger, *Nature*, 2010, **464**, 1218–1222.

35. G. F. Schröder, M. Levitt and A. T. Brünger, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2014, **70**, 2241–2255.
36. G. P. Poornam, A. Matsumoto, H. Ishida and S. Hayward, *Proteins*, 2009, **76**, 201–212.
37. O. S. Smart, T. O. Womack, C. Flensburg, P. Keller, W. Paciorek, A. Sharff, C. Vonrhein and G. Bricogne, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2012, **68**, 368–380.
38. G. N. Murshudov, P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long and A. A. Vagin, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2011, **67**, 355–367.
39. R. A. Nicholls, F. Long and G. N. Murshudov, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2012, **68**, 404–417.
40. R. P. Joosten, K. Joosten, G. N. Murshudov and A. Perrakis, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2012, **68**, 484–496.
41. P. A. Karplus and K. Diederichs, *Science*, 2012, **336**, 1030–1033.
42. V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson and D. C. Richardson, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2009, **66**, 12–21.
43. F. Dimaio, N. Echols, J. J. Headd, T. C. Terwilliger, P. D. Adams and D. Baker, *Nat. Methods*, 2013, **10**, 1102–1104.
44. R. I. Enchev, D. C. Scott, P. C. A. da Fonseca, A. Schreiber, J. K. Monda, B. A. Schulman, M. Peter and E. P. Morris, *Cell Rep.*, 2012, **2**, 616–627.
45. D. Turk, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2013, **69**, 1342–1357.
46. A. T. Brunger, *Structure*, 2005, **13**, 171–172.
47. G. G. Kotiguda, D. Weinberg, M. Dessau, C. Salvi, G. Serino, D. A. Chamovitz and J. A. Hirsch, *J. Biol. Chem.*, 2012, **287**, 42031–42041.

CHAPTER 8

Crystallography with X-ray Free Electron Lasers

N. A. ZATSEPIN[a,b]

[a] Arizona State University, Department of Physics, Mill Ave, Tempe 85287-1504, USA; [b] Arizona State University, Biodesign Center for Applied Structural Discovery, East Tyler St, Tempe, AZ 95287-7401, USA
Email: nadia.zatsepin@asu.edu

8.1 X-ray Free Electron Lasers – An Introduction

Historically, the invention of novel light sources has profoundly augmented our ability to probe and understand nature. The advent of hard X-ray free electron lasers (XFELs) was no exception. The distinction of XFELs lies in their production of multi-gigawatt, collimated, spatially coherent, femtosecond-duration X-ray pulses. Providing 9–10 orders of magnitude higher peak brightness than existing 3rd generation synchrotrons established the XFEL as a compelling source for the study of micron and submicron-sized samples and ultrafast phenomena. Regimes previously inaccessible at synchrotrons, or longer wavelength FELs, opened up opportunities in scattering and spectroscopic techniques in atomic, molecular and optical physics, femtochemistry, dynamical processes in nanoparticles, the creation and study of plasmas in the warm dense matter regime, and structural biology, particularly through serial femtosecond crystallography (SFX).[1] The first 5 years of hard X-ray FEL experiments are reviewed in detail by Bostedt *et al.* (2016), Schlichting (2015) and Spence (2017).[1–3]

Unlike synchrotrons, which use an electron storage ring to deliver beam to many experimental stations concurrently, XFELs are based on a long,

Chemical Biology No. 8
Protein Crystallography: Challenges and Practical Solutions
Edited by Konstantinos Beis and Gwyndaf Evans
© The Royal Society of Chemistry 2018
Published by the Royal Society of Chemistry, www.rsc.org

single-pass undulator. Due to very low normal-incidence reflectivity of mirrors for vacuum/extreme ultraviolet UV (VUV) and X-rays, VUV and X-ray FELs cannot use mirrors for a multi-pass setup for light amplification, as is possible for infrared, visible and UV FELs. In an XFEL, the wavelength of the fundamental on-axis undulator radiation, λ_r, that relativistic electrons with energy $E = \gamma m_e c^2$ emit as they traverse the undulator, forced into a sinusoidal trajectory (Figure 8.1A), is well approximated by eqn (8.1)

$$\lambda_r \approx \frac{\lambda_u}{2\gamma^2}\left(1 + \frac{K^2}{2}\right) = \frac{2\pi c}{\omega_r} \tag{8.1}$$

where the undulator strength parameter $K = eB_0\lambda_u/(2\pi m_e c)$, e is the electron charge, B_0 is the peak magnetic field strength, λ_u is the undulator period, $\gamma = E/(m_e c)^2$, m_e is the rest mass of the electron, c is the speed of light in vacuum, ω_r is the fundamental undulator frequency (see *e.g.* ref. 4 for details and a history of XFEL development). Higher harmonics are also generated but are much weaker. X-rays with energies of up to ~13 keV (25 keV in some cases) are available. The resonant interaction between the electron bunches and emitted radiation forces electrons into microbunches separated by the fundamental wavelength λ_r. Electron microbunching is enhanced by the growing radiation field, in turn increasing their radiative output, so the radiation field increases exponentially along the undulator until saturation, as depicted in Figure 8.1B. Electrons within a single microbunch emit radiation coherently, and thus the initial spontaneous radiation is amplified until a single mode remains. Since FEL electrons are not bound to atoms and thus not limited to particular energy transitions, the FEL wavelength is widely tunable. The radiation is fully polarized and transversely coherent. However, the longitudinal coherence of self-amplified spontaneous emission (SASE) XFELs is relatively poor ($\Delta E/E \sim 0.1\%$), and the SASE pulse spectrum is spiky and varies from shot to shot, as does the intensity, as a result of the shot-noise startup. Additionally, the SASE mode requires extremely bright electron beams and very high quality undulator arrays.[1]

The first hard X-ray FEL, the Linac Coherent Light Source (LCLS) at SLAC National Accelerator Laboratory near Stanford University, California, first lased in 2009. There are only a few operational hard X-ray FELs in the world: LCLS, the SPring-8 Ångstrom Compact free electron LAser (SACLA) in Japan, which came online in 2011, PAL-XFEL in Pokang, South Korea, and the European XFEL in Hamburg, Germany (EuXFEL), which both saw first light in 2017, and SwissFEL in Villigen, Switzerland, planned for commissioning in 2018. The linear design of an XFEL limits the number of experiments carried out at a time to one, or at best a few. Multiplexing of various forms has been used, for example, by diffracting a narrow part of the SASE spectrum to another experimental endstation, or by refocusing the so-called spent beam downstream of the weakly interacting sample, passed through a hole in the detector.[5] The growth in available XFEL beamtime is thus relatively slow and remains very competitive. XFEL crystallography experiments

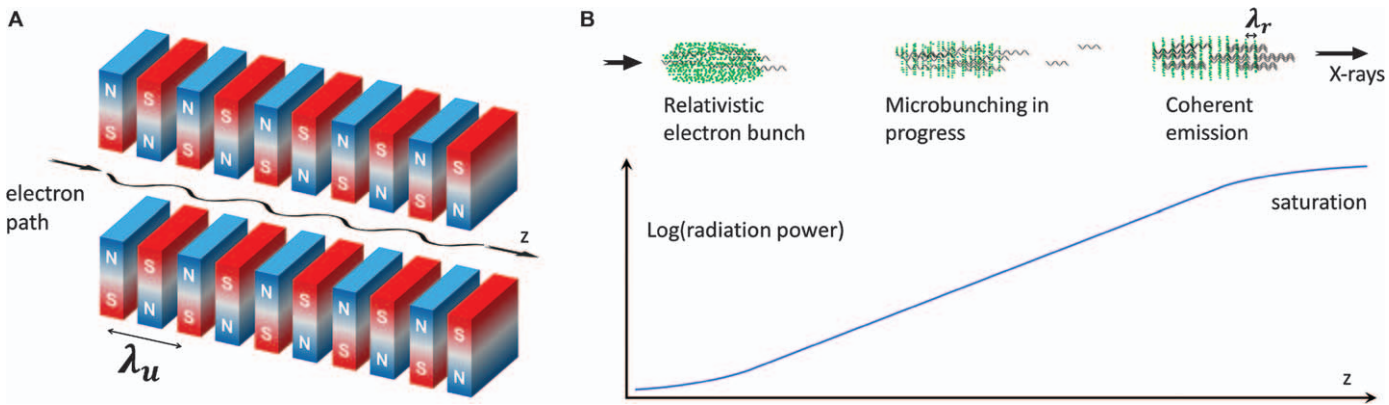


Figure 8.1 Schematic of self-amplified spontaneous emission of undulator radiation responsible for the extremely bright XFEL pulses. (A) The sinusoidal path electrons follow as they traverse the undulating magnetic field. (B) Correlated emission of X-rays from electrons in microbunches separated by one (fundamental) wavelength, which leads to exponentially increasing radiation emission until saturation.

require broad expertise, often involve large interdisciplinary collaborations, take months to plan and prepare, and are followed by months of data analysis. A lot of effort is being invested in method development (sample characterization, delivery and data analysis) to require less beamtime and smaller data volumes, and to extract more information in novel ways from the often-scarce samples, especially in structural biology. These developments are discussed below.

8.2 Radiation Damage at XFELs

Resolution in biological imaging and crystallography is typically limited by radiation damage,[6] which is particularly problematic for small crystals and metalloproteins. In 2000 it was suggested that using very high X-ray dose rates and ultrashort exposures may enable the collection of useful structural information before radiation damage destroys the sample.[7]

Consider the photoelectric cross-section of carbon for X-rays of 1 Å, which is about 10 times higher than its elastic cross-section. A molecule of T4 lysozyme (with 118 bound solvent molecules) experiences about 2000 primary ionization events from a flux of 3×10^{12} 12 keV X-ray photons in a 100 nm diameter beam.[8] In lighter elements (C, N, O and S), where Auger decay dominates over fluorescence, most photoelectric events remove two electrons from the elements, of which roughly 1 in 5 deposit additional energy into the molecule. High-energy photoelectrons can escape the sample if the sample is smaller than the mean free path of the photoelectrons. For samples larger than the photoelectron mean free path (about 130 Å for a 12 keV electron, rapidly dropping with energy), the cascade of electrons from subsequent collisions quickly thermalizes, resulting in sample temperatures reaching beyond 500 000 K.[9,10] For single particles, XFEL pulses as short as 10 fs were suggested to be required to obtain atomic resolution.[8] For crystals, however, spreading the dose over identical copies of the molecule decreases the dose a single molecule absorbs so longer exposures are tolerable.[11] Additionally, due to shielding of the ionized atoms by free electrons, the Coulomb explosion starts at the surface, so crystalline diffraction can still be obtained during the initial stages of the explosion.[12]

The radiation dose from XFEL pulses received by macromolecular crystals is multiple orders of magnitude higher than the critical damage dose for synchrotron crystallography of 0.2 MGy at room temperature (30 MGy for cryocooled crystals).[13] Critically, photoreduction of metal centers occurs at significantly lower doses than these limits.[14] However, by using pulses that terminate faster than the onset of structure-altering radiation damage, the room-temperature macromolecular structures determined with SFX are virtually damage free.

In the first SFX experiment, 2 keV X-ray pulses with 1×10^{13} photons pulse^{-1} deposited 700 MGy in submicrometer-sizes Photosystem I crystals.[15] At the observed resolution of 8.5 Å (limited by soft X-rays), no difference in scattering factors was observed between the datasets collected with 10 fs

and 70 fs pulses and it was estimated that about 700 MGy is the 'safe dose' for 70 fs XFEL pulses.[10] The crystals were also shown to withstand much longer pulses, with strong low-resolution Bragg spots visible from pulses up to 300 fs in duration.[15] High resolution detail was lost first, an observation that can be described as a self-terminating diffraction signal due to loss of crystallinity.[9] Higher doses, in shorter pulses, have also been used, with no detrimental effect on the quality of the determined structure. A 2 Å-resolution structure of native nanocrystalline granulovirus has been reported, where a dose of up to 1.3 GGy was delivered to $\sim 0.2 \times 0.2 \times 0.4\ \mu m^3$ crystals by 50 fs, 7.95 keV X-ray pulses.[16] These are the smallest protein crystals that yielded a high-resolution structure from SFX. Doses beyond 1 TGy are expected to be tolerable for pulses with sub-femtosecond duration.[17]

The majority of high resolution SFX data has been collected at the Coherent X-ray (CXI) endstation at LCLS, from 1–10 μm-sized crystals.[16] In the vacuum sample chambers at CXI, the beam is focused to 0.1 or ~1 μm FWHM by a pair of Kirkpatrick Baez mirrors, and energies of 5–12 keV are available. Although most SFX experiments at CXI used 40–50 fs hard X-ray pulses, it may be necessary to use even shorter pulses (20 fs is suggested) to reduce damage effects on high-Z atoms.[18] With such short pulses, simultaneous (femtosecond) spectroscopy and diffraction can be used to probe the structural and electronic state of redox-active catalytic centers in ambient conditions during reactions.[19–21]

Room-temperature (RT) crystallography has multiple advantages over cryo-crystallography, most importantly enabling time-resolved studies at unprecedented time scales in physiologically relevant environments. RT crystallography circumvents the search for suitable cryoprotectants. Additionally, cryo-crystallographic studies sample only a subset of conformations,[22] biasing structures towards smaller, overpacked models, potentially hiding conformational states that are key to physiological function.

While the dominant advantage of XFEL use is outrunning secondary radiation damage, by varying the fluence in SFX measurements, the induced changes in the scattering factors of heavy atoms could be used for direct phasing in a process called high-intensity radiation damage-induced phasing (HI-RIP). HI-RIP is analogous to synchrotron cryo-crystallographic RIP, which makes use of disulfide bond breakage and/or ionization of heavy metals[23] for substructure determination.

8.3 Serial Femtosecond Crystallography

8.3.1 SFX Experimental Setup

A hard X-ray FEL's brilliant, femtosecond-scale X-ray pulses enable the collection of high-resolution diffraction from macromolecular micron and submicron-sized crystals at room temperature, prior to radiation-damage-induced atomic motion, in a close-to-native environment. A single ~50 fs

XFEL pulse (with 10^{10} photons focused to a 1 μm-diameter beam) will destroy a micron-sized protein crystal, so only a single diffraction snapshot can be recorded from each crystal. Each diffraction snapshot is read out independently. Due to the brevity of the XFEL pulses, there is no need to freeze the sample to minimize radiation damage. The crystal is still during the exposure, but is subsequently destroyed by the Coulomb explosion resulting from its ionization by the brilliant XFEL pulse, earning the technique the name ''diffraction before destruction''.[24] Fresh sample needs to be brought into the interaction region before the next XFEL pulse, so microcrystals are rapidly delivered across the pulsed XFEL beam in a liquid or viscous jet (Figure 8.2), scanned across the beam on a fixed-target sample holder or delivered by other means as discussed in § 8.3.2. This data collection mode is called serial femtosecond crystallography (SFX) and the analogous serial crystallography at synchrotrons (using necessarily longer exposures) is referred to as serial millisecond crystallography (SMX) or serial synchrotron crystallography.

Diffraction snapshots from thousands of microcrystals in different orientations are merged to assemble the full 3-dimensional reflection list, which can then be phased by established crystallographic phasing protocols. The spatial coherence and brevity of XFEL pulses has additionally provided new opportunities in direct phasing, which will be discussed in § 8.5. The next sections will cover SFX highlights, various microcrystal

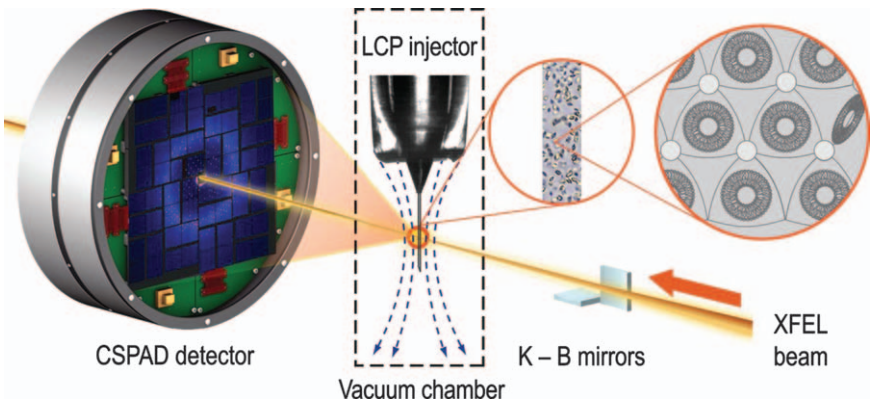


Figure 8.2 An example of an experimental setup for serial femtosecond crystallography with an XFEL using a lipidic cubic phase (LCP) injector (all under vacuum). Microcrystals of GPCRs in LCP can be delivered to the XFEL beam path in a continuous viscous column of ~20–50 μm diameter LCP. The XFEL beam is focused using Kirkpatrick-Baez (KB) mirrors, and data collected on the Cornel-SLAC Pixel Array Detector (CSPAD).[39] The insets represent a viscous crystal slurry, with the second inset showing a schematic of the structure of LCP.
Figure from Liu *et al.* Serial femtosecond crystallography of G protein-coupled receptors. *Science*, **342**, 2013. Reprinted with permission from AAAS.

sample delivery options for static and time-resolved SFX, data analysis and software, along with a number of exciting new directions being explored in XFEL crystallography that exploit the unique properties of XFEL pulses.

8.3.2 SFX Early Achievements

The first proof-of-principle 'diffraction before destruction' SFX experiment (at the Atomic and Molecular Optics beamline at LCLS[15] in late 2009) used soft X-ray pulses and membrane protein photosystem I nanocrystals as a model system. Since then, impressive developments in sample characterization, delivery, analysis, detectors and XFEL beamlines have helped evolve SFX into a successful method for room-temperature protein structure determination, pump-probe time-resolved crystallography and recently, mix-and-inject time-resolved crystallography. By outrunning atomic motions due to radiation damage, SFX has allowed researchers to determine structures from microcrystal slurries in cases where it could have taken years to optimize conditions for large, high-diffraction-quality crystals for use at synchrotron beamlines. The current XFELs' high brilliance allows high resolution diffraction patterns to be obtained from crystals with only 100 unit cells in volume (under ideal conditions).[16]

One of the key successes of SFX is structure determination of membrane proteins and large membrane complexes (see[25] for a review), including those responsible for photosynthesis: photosystem I,[15] light-activated states of photosystem II,[19,26–29] cytochrome c oxidase,[30,31] and photosynthetic reaction centers,[32] as well as a medically important family of eukaryotic membrane proteins that mediate cellular signal transduction, G-protein-coupled receptors (GPCRs).

GPCRs comprise about half of the current pharmaceutical drug targets, but until recently, GPCR structure determination was very slow. In recent years, protein engineering and the use of the lipidic cubic phase (LCP) for crystallization greatly facilitated and accelerated GPCR structure discovery.[33] LCP is a liquid crystalline phase that mimics the lipid bilayer in the cellular membrane, where GPCRs are found. Using LCP, or the *in meso* method, is a robust approach to crystallization of membrane proteins that led to a massive boost in new membrane protein structure determination in the last decade.[34] The 2012 Nobel prize was awarded to Robert J. Lefkowitz and Brian K. Kobilka for studies of G-protein-coupled receptors, some of which were crystallized in LCP,[35,36] which contributed to the popularity of LCP use. Between the start of 2012 and September 2014, about 100 integral membrane proteins and peptides were solved from crystals grown *in meso* (out of a total of about 200 at that time, since 1996).[34] Despite LCP facilitating the formation of high-quality GPCR crystals, they are often small (sub-10 μm) crystals[37] and thus particularly well suited candidates for SFX. The invention of a microfluidic injector that delivers protein crystals to the XFEL beam in LCP established LCP-SFX as a reliable approach to GPCR structure

determination.[38] By directly injecting the tiny, fragile membrane protein microcrystals in their growth medium into the XFEL path, microcrystal loss or damage through manual handling is minimized. Another critical advantage of LCP and other highly viscous jets is significantly lower consumption of sample (nanograms of protein) compared to the fast-flowing liquid jets. Membrane protein structures determined by LCP-SFX include the human smoothened (SMO) receptor in complex with cyclopamine,[38] the human serotonin receptor,[39] diacylglycerol kinase,[40] rhodopsin (involved in human vision) bound to arrestin,[41] an angiotensin receptor that could only be imaged with its bound ligand drug molecule using LCP-SFX (as shown in Figure 8.3[42,43]), and a δ-opioid receptor, involved in pain relief.[44] Alternative high-viscosity media have also been used for SFX crystal delivery in the high-pressure LCP injector or similar devices, which are discussed in the next section.

The ability to use SFX to determine structures from crystals grown in cells, without their extraction (*in vivo* crystallography) has also been demonstrated,[45] potentially enabling crystallization of proteins and macromolecules that are particularly difficult to extract and purify.[46]

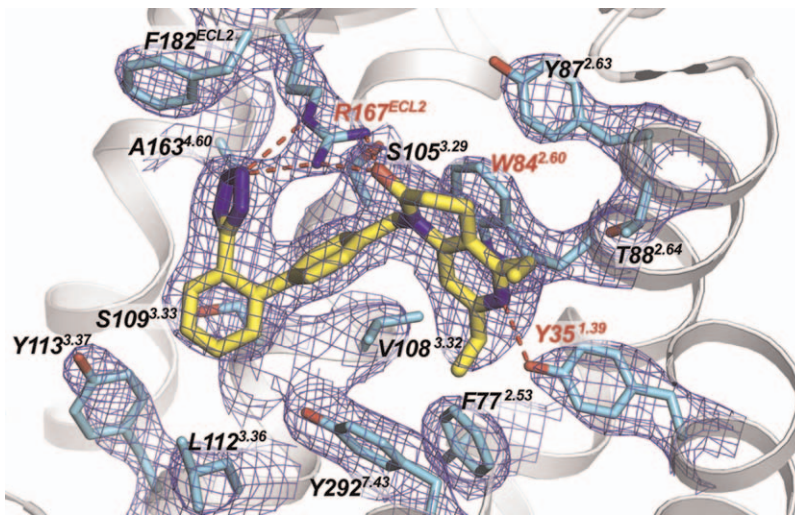


Figure 8.3 Example of 2.9 Å-resolution electron density of a human GPCR that is involved in blood pressure regulation, the angiotensin II type 1 receptor, with a bound ligand (antagonist ZD7155). Identifying critical structural and binding interaction details was made possible by prominent, well-defined electron density in the ligand-binding pocket obtained from high quality, room temperature LCP-SFX data. This 2mFo-DFc map (blue mesh) contoured at 1σ, was calculated from 73 130 indexed diffraction patterns.[42]
Reprinted from Cell, 161, Zhang, H. *et al.* Structure of the Angiotensin Receptor Revealed by Serial Femtosecond Crystallography, 833–844, 2015, with permission from Elsevier.

8.3.3 SFX Sample Delivery and Data Collection Rates

The delivery of a new crystal to the XFEL interaction region needs to be efficient and reliable. Ideally, each XFEL pulse hits a new crystal and no crystals are wasted between shots. At LCLS, XFEL pulses of ~50 fs duration arrive at 120 Hz so a suitable solution to rapidly replenishing the microcrystal, minimal background X-ray scattering, while keeping protein crystals at room temperature and in their physiologically relevant environment, is to spray them into the path of the XFEL with a liquid injector.[47,48] The most often-used sample delivery for SFX is a liquid jet with a gas-dynamic virtual nozzle (GDVN) that uses concentric gas flow to focus the crystal-laden stream to a ~1–6 μm-diameter stream[47,49] (akin to that shown in Figure 8.2). The gas focusing creates a very thin jet (and thus low background scattering) without the need for a small nozzle exit hole, which can clog. It is a sample delivery system that maintains the crystals in a fully solvated state and avoids sample ionization during delivery, undue pressure, temperature or shear on the crystals.[48] However, a liquid jet flowing at 10 m s^{-1} advances the microcrystals a couple of centimeters between the very brief, micro-scale XFEL pulses, so the majority of protein crystal solution is wasted between XFEL shots. With these parameters, which are representative of the first years of SFX experiments, only about 1 out of 10 000 microcrystals is probed by the 0.1–2 μm diameter XFEL beam.[38]

Millions of detector images are collected during each experiment, but the number of useful frames, that is those that contain potentially indexable diffraction spots from a crystal, is much smaller. The 'hit rate', or the ratio of the number of crystals intercepted by the XFEL pulses to the number of XFEL shots read out from the detector, fluctuates significantly. For a random distribution of crystals in solution delivered with a liquid injector, the ideal hit rate to maximize single crystal hits is ~63% (based on Poisson statistics), of which just over half should be single crystal hits.[50] However, at the necessary concentrations, crystals can clump and block the microfluidic injectors, so microcrystal slurries are diluted to facilitate jetting. From a practical perspective, a pattern is considered a crystal hit if it contains at least 15 peaks (this number is flexible and is guided by autoindexing software capabilities). For most samples in the first few years of GDVN-based SFX, hit rates of 0.1 – 10% were typical. The reasons for low hit rates include, but are not limited to, injector nozzles clogging from dehydration of the buffer (in vacuum) on the surface of the nozzle exit hole, or the micron-scale jet flickering out of the path of the micron-scale XFEL beam path, or the microcrystals may simply be too small or weakly diffracting. Additionally, early SFX indexing software was unable to index multiple crystals per pattern so that also influenced decisions about dilution of crystal slurries for data collection. Realistically, consistent hit rates above 10% are considered good. At the LCLS frame rate of 120 Hz, a 10% hit rate and can yield ~35 000 indexed patterns from 1 hour of data collection, assuming an indexing rate of 80%. This can be enough to solve a structure (using molecular replacement)

if the patterns are of high quality. The number of patterns necessary to determine a structure or answer a specific biochemical question is highly dependent on sample and experimental conditions and fluctuates significantly. It has, however, dropped by orders of magnitude due to significant improvements in all aspects of SFX technology, including detectors, sample delivery, detector metrology and, critically, intensity scaling and post-refinement. This is discussed in more detail in § 8.5.

While GDVNs were used for structure determination for a number of years (*e.g.* ref. 15, 51–53), the level of sample consumption with the early GDVNs (roughly 10 μL minute^{-1}, requiring 10–100 mg of pure protein, with most of it wasted) precluded the use of SFX for particularly precious, scarce samples. A number of approaches have been developed to mitigate this, including thinner GDVN jets, a slow-flowing high-viscosity injector,[38] raster-scanning fixed-targets (*e.g.* ref. 50, 54–58), electro-kinetic injection[59] or acoustic droplet ejection.[21] A recently developed double-flow-focused liquid injector provides a stable ultra-thin liquid jet and thus lower background scattering, a decrease in sample consumption and improved jet stability.[60]

The electrokinetic injection ('electrospinning') involves emitting a continuous filament (that becomes a solid thread due to rapid dehydration) and significantly reduces sample waste compared to GDVN with a flow rate of ~0.17 μl minute^{-1}.[59] Unlike electrospray, in electrospinning, microjet droplet formation is prevented by the addition of antifreeze (*e.g.* glycerol), limiting its use to samples compatible with such additives.

Acoustic droplet ejection (ADE) coupled with a conveyor belt drive, dubbed 'drop-on-demand' sample delivery, lowers sample consumption (compared to GDVNs) and allows multimodal data collection – simultaneously collecting SFX and X-ray emission spectroscopy for time-resolved studies of metalloproteins.[21]

A particularly effective way to increase crystal hit rate is to use a fixed target sample holder that is raster scanned across the XFEL beam, optimized for micron-sized crystals, with low background scattering, and prevents severe dehydration or temperature changes. To try to address this need, a number of fixed target sample holder systems have been developed for XFEL crystallography.[50,54,55,58] A microfabricated crystalline silicon chip with micropores,[55] on a translatable stage called the Roadrunner goniometer recently enabled crystal structure determination of the picornavirus bovine enterovirus 2 and the cytoplasmic polyhedrosis virus type 18 polyhedrin, from impressively brief data collection: only 14 and 10 minutes, respectively, at LCLS (at 120 Hz).[61] The silicon chip contributes almost no background scattering and can be used in air and with a humidified helium stream. With enough micropores to hold a few thousand micro/nanocrystals (only 1–3 μl of crystalline suspension), protein and beamtime are used much more efficiently. Additionally, the buffer can be wicked off and the sample flash frozen by plunging into liquid nitrogen for serial cryo-crystallography at a synchrotron, potentially without the need for cryoprotectants for crystals smaller than 20 μm.[61]

Fixed targets were used measure diffraction patterns from individual single-layer 2D crystals of a membrane protein at room temperature, for the first time[62,63] with resolution reaching 4 Å,[64] and could potentially be used for time-resolved SFX on 2D crystals. Another fixed-target system is comprised of microfluidic crystal 'trap array' chips in which microcrystals are flown across the chip and get trapped in fixed, addressable positions, and can then be raster scanned across the beam to maximize crystal hit rates.[54] Hit rates from crystals deposited on fixed targets can reach over 90%, and for high quality samples, with optimized detector geometries and SFX software now capable of multi-crystal indexing (per pattern), indexing rates >100% are increasingly common. These numbers are optimistic and sample distribution in liquid injection, or even on fixed targets, is non-uniform; and, naturally, indexing quality fluctuates from crystal to crystal.

One of the biggest breakthroughs in the early days of SFX was the development of the high-pressure injector mentioned earlier, designed for use with lipidic cubic phase (LCP)[38] or other high-viscosity media. Due to the much slower flow rates, protein volumes required for structure determination from LCP-SFX are at least an order of magnitude smaller than with GDVN liquid jet delivery. The LCP injector comprises a hydraulic stage driven by a high-performance liquid chromatography (HPLC) instrument, a sample reservoir and a nozzle. Nozzles of 10–50 μm inner diameter are typically used, and coaxially flowing gas helps maintain continuous on-axis flow[65] (see Figure 8.2). To avoid LCP jet breakup (and slow recovery due to its viscosity and slow flow rate), it is best practice to attenuate the XFEL beam to just below the LCP LCP-breaking intensity (found manually during experiments). Hit rates of at least 10% are not uncommon in LCP-SFX experiments.

For macromolecular crystals that are incompatible with LCP, various viscous media are being explored for post-crystallization mixing and gentle, efficient sample delivery, such as agarose,[66] high-molecular-weight poly(ethylene oxide) (PEO, with molecular weight 8 000 000),[67] and a grease matrix carrier.[68] Using the agarose jet (which is mostly water), for example, also mitigates one disadvantage of LCP, which is the presence of strong background scattering at low resolution (>30 Å) and a strong diffuse ring around 4.5 Å.

For most time-resolved experiments involving a pump probe laser (discussed later), standard GDVN liquid jets have been used (*e.g.* ref. 69). The use of silicon chip-based fixed targets for pump-probe SFX has been also been explored.[58,61] Fixed targets can provide more accurate control of time delays including much longer delays than is practical for a microfluidic nozzle. Pump-probe serial crystallography (PP-SFX) has also been demonstrated in the LCP injector, using microcrystals of bacteriorhodopsin.[70] The accumulation of M intermediate by time-resolved visible absorption spectroscopy supported the difference Fourier maps observed at 1 ms delay, making these results particularly promising as they combine the very sample-efficient delivery of LCP-SFX with the advantages of PP-SFX.

For time-resolved enzymology experiments, a number of 'mixing jets' are being developed for enzymology studies or 'mix-and-inject' serial crystallography (MISC). The first two reported MISC experiments implemented similar T-junction mixers which provided a 10 second mixing time for ligand binding to the aptamer domain of a riboswitch,[71] and 2 second mixing time for the reaction of *Mycobacterium tuberculosis* ß-lactamase microcrystals with ceftriaxone antibiotic solution.[72,73] Improved mixing injectors will enable faster mixing times to provide shorter time delays. One example was a proof-of-principle mixing injector with coaxial supply lines for reactant and crystal solution, which mix prior to injection into the XFEL beam path, potentially achieving submillisecond mixing times with an adjustable delay between mixing and XFEL probing, but at the cost of a high sample dilution ratio.[74] An improved design for a mixing injector has been demonstrated that improves the hit rate and provides better control over mixing times.[75] The latter can comfortably reach submillisecond mixing times, and mix-then-probe delay times from submillisecond to hundreds of milliseconds are directly accessible by fabricating devices with varying delay lengths.[75]

The ideal microcrystal delivery system for SFX depends on (a) the sample, including crystal size distribution and number density in solution; (b) whether the crystals can maintain diffraction quality after sitting in vacuum or in a humidified He stream on a patterned silicon chip; (c) crystal sensitivity to temperature fluctuations; (d) buffer compatibility with liquid or high-viscosity injector and nozzles and vacuum or in-air operation, as well as on (e) the XFEL beam size and timing structure. To minimize protein waste, microcrystal delivery should be matched to the source repetition rate, replenishing the sample while also moving it sufficiently downstream to avoid exposing a damaged part of the sample to subsequent XFEL pulses. Warm linac-based XFELs with low repetition rates (up to 60 Hz at SACLA and PAL-XFEL, 120 Hz at LCLS, 100 Hz at SwissFEL) are better suited for use with slower sample delivery. The high pulse rates at recently commissioned superconducting linacs, (up to 4.5 MHz planned at the European XFEL and up to 1 MHz planned at LCLS II) can make efficient use of existing or, preferably, higher GDVN flow rates. Assuming the damaged region radius in a liquid jet is about 15 μm (estimates vary and depend on medium), a jet would need to move $\sim$70 m s^{-1} to move the crystal 15 μm before the next XFEL pulse at the planned pulse spacing of 220 ns at the European XFEL.

8.4 Time-resolved Serial Femtosecond Crystallography

In their functional state, proteins are not static like the thousands of lattice-constrained, spatially and temporally-averaged structures determined from cryo-crystallography may imply. It is their dynamic nature and their transient states following photon absorption, molecular binding, temperature or pH changes that play a critical role in the behavior and development of many

living systems, whether by driving electron transfer across a membrane or isomerization or formation of a covalent bond. SFX has enabled room-temperature dynamic studies of macromolecules using light-activated and ligand-binding reaction initiation.

8.4.1 Pump Probe Serial Femtosecond Crystallography

Photoreceptors, *i.e.* the combination of a protein and chromophore (which absorbs photons in the visible range) are responsible for photosynthesis, phototropism, photomorphogenesis in higher plants, image formation and circadian rhythm regulation in mammalian eyes, and phototaxis in bacteria.[76] While biologically relevant processes have time scales ranging from microseconds to much longer, the initial chemical reaction after photon absorption is of the order of femtoseconds and requires ultrashort pump and probe pulses.

Pump-probe crystallography, typically using the Laue method at synchrotrons, is an established method where the reaction initiation (pump) is followed by an X-ray probe or series of probes. It is a powerful approach to determining protein kinetics and structures of intermediate states. However, radiation damage has limited time-resolved studies at synchrotron sources to well-behaved and, typically, reversible conformational changes, and requires the growth of sufficiently large, high quality crystals.[77] XFEL pump-probe SFX (PP-SFX) exploits ultrafast (fs–ns) pump lasers and crystals smaller than the absorption length of the pump laser to synchronize reaction initiation in the crystal. Using larger crystals, as necessary at a synchrotron, means only 10–20% of molecules in the crystal undergo reaction initiation.[78]

In PP-SFX, microcrystals are exposed to the reaction-initiating pump laser upstream of the XFEL interaction region and time delays can be controlled by the liquid flow rates and distance between pump and probe positions on the liquid jet. The time resolution is then a convolution of pump and probe laser profiles and timing system jitter.[78] Each frame of the time series requires thousands of crystals in random orientations, and the accuracy of merged intensities needs to be very high to observe the small structure amplitude differences between light and dark structures. Current implementations of the diffract-and-destroy method treat reversible and irreversible reactions equivalently as each crystal is destroyed after being hit by the ionizing XFEL pulses.

One of the most significant achievements of XFEL SFX is atomic-resolution time-resolved crystallography on time scales (fs) inaccessible at synchrotrons, which are limited to about 100 ps by their pulse structure. Time-resolved SFX was first demonstrated in 2012, on the large membrane protein photosystem I-ferredoxin cocrystals, with structure factor changes observed around a pump-probe delay of 10 μs.[79] Time-resolved structures that have been determined with pump-probe SFX include *trans–cis* isomerization of the chromophore in photoactive yellow protein over a time

range from 100 fs to 3 ps,[69,80] photoactivation-induced conformational changes in bacteriorhodopsin from nanoseconds to milliseconds,[81] light-activated states of photosystem II, a large membrane protein complex involved in water oxidation/oxygen production in plants and algae,[19,26–28,82] photolysis of carbonmonoxy myoglobin over a range of time delays between 0.1 ps and 150 ps,[83] and chromophore twisting in the excited state of a reversibly photoswitchable fluorescent protein.[84]

Using light is the fastest way to trigger a synchronized reaction. Proteins that are not photoactive can be made light sensitive by genetic engineering (optogenetics) or using photoactivatable bioagents in the form of caged molecules, making many more targets amenable to pump-probe SFX.[77]

8.4.2 Mix-and-inject Serial Femtosecond Crystallography

To study enzymatic reactions, an initially inactive substrate present in the crystal can be activated by diffusion of a small molecule into the crystal, which is then probed by X-rays. Mix-and-inject serial crystallography (MISC) exploits the rapid mixing and fast diffusion rates possible with tiny crystals, at room temperature.[73,85,86] For the size of crystals typically studied at synchrotrons (several tens of micrometers), spatial equilibration by diffusion requires hundreds of milliseconds, compared to a few microseconds for the 1–5 μm crystals typical in SFX experiments.[76]

The shortest time point that can be probed with MISC depends on mixing and diffusion rates. Using tiny crystals (with large surface to volume ratio) and current microfluidic injection systems, designed to optimize rapid mixing of crystal and substrate solutions, has enabled time points of the order of seconds to be probed.[71,72] Reactions that cannot readily be initiated in the crystal may be initiated in solution and studied with pump-probe small and wide angle X-ray scattering (SAXS and WAXS) at XFELs.[87,88]

Pump probe and mix-and-inject XFEL experiments are not limited to crystals. Other reaction activation approaches may also be possible, such as pH or temperature jumps, or application of voltage gradients (for 2D crystals), or light-triggered ligand release with pre-soaked crystals.

8.5 Serial Femtosecond Crystallography Data Analysis

8.5.1 SFX Data Collection Overview

The ultrashort XFEL pulses record snapshot diffraction patterns from randomly-oriented crystals that are essentially still during the brief exposure. Thousands of crystals are needed for accurate reflection intensities as almost all reflections are partially recorded. The large data volumes, rapid data collection rates, and the challenges involved in accurately merging data from big numbers of individual crystals collected with stochastically fluctuating

XFEL pulses necessitated the development of new protocols and software for data collection and analysis.

SFX data analysis will be considered here in 4 parts: §8.5.2 data collection monitoring, §8.5.3 data reduction and hit finding, §8.5.4 Bragg diffraction analysis: high-throughput, parallelized autoindexing, merging and post-refinement (Figure 8.4), and §8.5.5 phasing and model refinement. In §8.5.6 we will cover the achievements in *de novo* phasing of SFX data. We will also discuss novel phasing approaches to SFX data that exploit the coherence and brevity of XFEL pulses. Software described in this section has mostly been developed by groups outside of XFEL facilities and are mostly facility independent. Where facility-specific details are important, we focus on LCLS where the majority of SFX experiments have been carried out.

8.5.2 Data Collection Monitoring

Sample and experimental time are precious, rendering real-time monitoring of quantity and quantity of XFEL data collection crucial for rapid feedback and decision making during beamtimes. The primary quantities monitored during SFX data collection are: a running average of the hit rate (diffraction patterns/number of detector frames recorded), diffraction resolution, numbers of peaks per pattern and number of saturated peaks. Provided accurate detector metrology, live indexing rates can also be monitored. Users can thus make quick decisions regarding when a sufficient number of indexable patterns has been recorded or whether to switch to different crystallization batches based on inadequate resolution or pattern quality.

A number of software packages have been developed for this purpose: CASS,[89] *psana*,[90] *Cheetah*,[91] primarily used offline for data reduction (step 2, below) but also available for real-time analysis at LCLS using *psana*, *cctbx.xfel* using the *Data Exploration Toolkit*,[92,93] *Hummingbird*[94] and *OnDA.*[95] Most of these are open-source and are either written in Python or include a Python interface. *Psana* (*Photon Science Analysis*) is developed at LCLS and although primarily written in C++, it provides user interfaces in C++ and Python. The trend at light sources around the globe is to move towards using Python, increasingly making more of these tools facility independent and more user friendly.

8.5.3 Hit Finding

SFX data collection rates are high: $\sim$2 TB h^{-1} at LCLS with 120 Hz pulses using the 14-bit 2.3 megapixel Cornell-SLAC pixel array detector (CSPAD[96]), and much higher at the superconducting linac XFELS where 1–4.5 MHz repetition rates are planned. As detailed in § 8.3.3, not every XFEL pulse hits a crystal. The first step in data reduction is to sort the detector frames into those containing crystal diffraction (hits) and those without (non hits). The obvious way to approach this is to search for frames which contain at least N peaks (candidate Bragg reflections) with some number of connected pixels

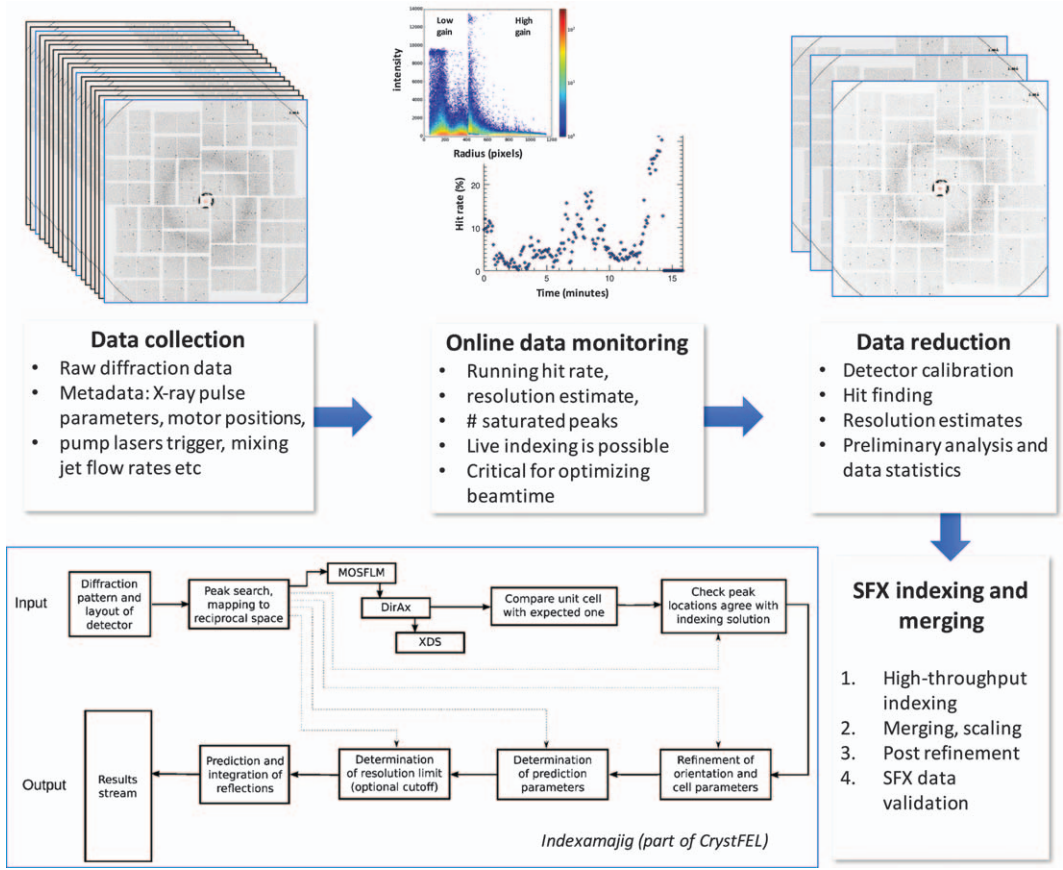


Figure 8.4 SFX data analysis from rapid data collection, monitoring hit rates and data quality, hit finding and detector calibration, high-throughput autoindexing, and a schematic of serial crystallography indexing program *indexamajig*, part of the *CrystFEL* software suite.[100]
The *indexamajig* flowchart is reproduced from White, T. A. *et al.* Recent developments in CrystFEL, *Journal of Applied Crystallography* **49**(2), 2016, licensed under CC-BY 2.0.

above a certain intensity threshold. Typically, autoindexing algorithms can handle $N \geq 15$, though, recently, novel sparse pattern indexing approaches are being developed to extract more from the raw data and index patterns with very few, weak peaks.[97]

This step of data reduction includes applying detector corrections like offset (dark current) subtraction, gain calibration, and masking of hot/dead/otherwise inaccurate pixels. Various additional masks can be applied to remove artifacts such as aperture shadows, strong scattering from the liquid jet or diffraction rings from a fixed target substrate or ceramic nozzles. *Cheetah* (https://stash.desy.de/projects/cheetah) and *cctbx.xfel* (http://viper.lbl.gov/cctbx.xfel/) are the most commonly used programs for SFX hit finding.

The large data volumes involved mean data storage and curation, especially for the older SFX experiments, was a considerable burden. Dedicated high-performance computing clusters are critical for efficient high-throughput data reduction and indexing (although the latter can be performed on an average laptop once the data are clean, but much slower). Reductions in data volumes, which could be tens of TB's after a single experiment, from improvements in analysis software have lessened computing and storage needs. However, computation that involves thousands of images simultaneously in memory (*e.g.* for accurate post-refinement) is very memory and computationally intensive. LCLS and the European XFEL fortunately have access to pre-existing computing infrastructure from the facilities' earlier focus on high energy physics. Despite this, data collection strategies will need to be adapted for the European XFEL's 4.5 MHz rates and for the planned superconducting source LCLS II (~1 MHz). The European XFEL can deliver 10 pulse trains per second, with up to ~350 XFEL pulses per train, in which pulses are separated by at least 220 ns. This very high rate of data collection will very likely require online hit finding and vetoing. Multiple new detectors were built specifically to handle the extreme data rates, such as the Adaptive Gain Integrating Pixel Detector (AGIPD[98]) installed at the Single Particles, Clusters, and Biomolecules & Serial Femtosecond Crystallography beamline (SPB/SFX).[99] AGIPD can handle a 5 MHz repetition rate (with frames read out in the 99.4 ms between pulse trains) and, due to its automatically adapted gain (3 settings), it can cover a dynamic range of 4×10^4 coincident photons.

8.5.4 Bragg Diffraction Analysis: Indexing, Merging, Post-refinement

Candidate crystal diffraction patterns are passed to a high-throughput autoindexing program which attempts to index each pattern independently. Since each crystal is hit only once (except for larger, slowly-flowing crystals), and due to the stochastic nature of the XFEL pulse, each diffraction pattern is effectively an independent measurement and so this autoindexing step is highly parallelizable. The final reflection list is obtained by merging

symmetry-related reflections from all indexed patterns. Post-refinement is discussed in detail in a later section.

CrystFEL is a suite of programs for processing serial crystallography data, including autoindexing, plotting unit cell distributions, merging partial reflection intensities from thousands of crystals, post-refinement, calculating data quality metrics, detector geometry optimization, and serial crystallography pattern simulations.[100–102] *CrystFEL* is available from www.desy.de/~twhite/crystfel/.

The data processing pipeline for SFX at SACLA was developed based on *Cheetah* and *CrystFEL*, adapted for data acquisition through the SACLA application programing interface.[103] It provides users with live data quality metrics including hit rates, resolution estimates, and indexing rates and unit cell distributions. The live plot of cumulative unit cell histograms (and resolution statistics) is an efficient way to see if good data are being collected and if there is contamination of the sample delivery lines by, *e.g.*, the calibration sample.

The *Data Exploration Toolkit* provides a way to explore SFX datasets' crystal heterogeneities prior to merging all diffraction data for a final reflection list.[93]Despite identical crystallization conditions, crystals within a single batch can differ significantly in diffraction quality or even unit cell size. Clustering large datasets makes better use of the measured data as reflection lists converge faster from more similar crystal distributions.[104]

The problem of merging Bragg diffraction patterns from thousands of microcrystals in random orientations was formulated by Kirian *et al.* (2010).[105] The fine angular collimation and 0.1% energy spread of the SASE XFEL pulse probes only a thin slice of the full angular profile of each Bragg reflection (exaggerated for clarity in Figure 8.5), which is broadened by nanocrystal size effects and mosaicity (for example in Figure 8.6). Each rotated mosaic 'block' traces out a hemispherical cap at the end of every reciprocal lattice vector (in 3D, centered at the origin) whose width increases with scattering angle, while strains within a crystal (variations in unit cell dimensions) also broaden the spots increasingly with scattering angle.[106,107] The angular deviation from the exact Bragg condition allows a degree of partiality, p, (between zero and unity) to be defined, according to eqn (8.2)[108]

$$\rho = \int_0^1 E(p)P(p)\mathrm{d}p \tag{8.2}$$

Here p is a coordinate along the arc of diffracted intensity on one diffraction pattern from a mosaic crystal, E describes the distribution of wavelengths in the incident beam and P is the distribution of scattered intensity along this arc. The partial reflections must be merged to form full reflections in order to analyze the resulting average structure factors.

In conventional protein crystallography, a goniometer is used to provide multiple accurate measurements from individual crystals. During the ultrashort XFEL pulses, the crystals are still, and the partially-recorded Bragg

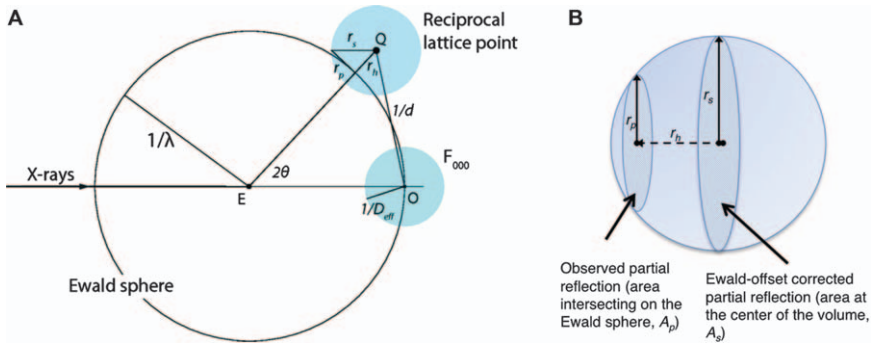


Figure 8.5 (A) Geometric definition of partiality. For a still shot taken with monochromatic X-rays of wavelength λ, a reciprocal lattice point (blue ball centered on Q) partially intersects the Ewald sphere. The intersection area, a spherical cap, is approximated by a circle of radius r_p, which is determined by r_h the distance from Q to the Ewald sphere, and r_s, the resolution-dependent radius of the reciprocal lattice point Partiality is defined as the intersection area-to-ball volume ratio for lattice point Q, normalized by the intersection area-to-ball volume ratio of the F_{000} spot at reciprocal space origin O. Figure reproduced from Sauter, N. K. XFEL diffraction: developing processing methods to optimize data quality. *J. Synchrotron Radiation*, **22**, 239–248, 2015, licensed under CC-BY 2.0 and with permission from the author. (B) Volume of Bragg spot with radius r_s with an Ewald offset correction of r_h. Figure reproduced from Uervirojnangkoorn, M. *et al.* Enabling X-ray free electron laser crystallography for challenging biological systems from a limited number of crystals, *eLife* 2015;**4**:e05421, licensed under CC-BY 4.0.

intensities reflect stochastic fluctuations in the crystals' random orientations, size, unit cell (possibly), structural heterogeneity and diffraction quality, the fluctuating SASE spectrum and intensity, and the unknown, random position of the particles with respect to the beam. The intensity 'samples' corresponding to a given Bragg reflection thus vary by orders of magnitude and the error in structure factor measurement from Monte Carlo integration over all of these random, mostly independent parameters, is proportional to $1/\sqrt{N}$ where N is the number of merged diffraction patterns (*i.e.* the number of nano/microcrystals used in the final dataset).[109] Simulations indicate that using the two-color modality (two narrow bandwidth XFEL pulses of slightly different mean energies arriving at the sample simultaneously), the error is still proportional to $1/\sqrt{N}$ but with a much smaller pre-factor.[110] Monte Carlo integration relies on having enough shots to provide sufficient random sampling across the angular profile of the Bragg peak to estimate its volume in reciprocal space. Simulations also indicate that using a slightly divergent or slightly polychromatic beam (to increase the partiality), or narrower crystal size distribution, can also expedite structure factor convergence.[101,104,111]

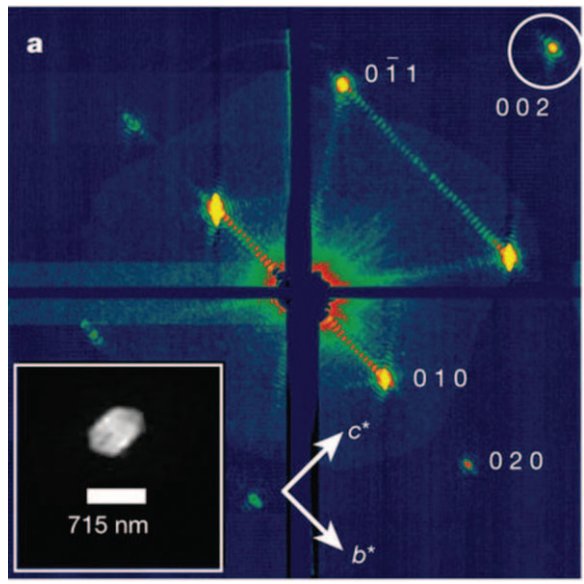


Figure 8.6 Scattering between Bragg reflections samples the diffraction pattern of a unit cell, which enables direct phasing of nanoparticles smaller than the beam. This figure (with intensities on a logarithmic scale) shows clear fringes between low-order Bragg reflections from a nanocrystal of a large membrane protein, photosystem I.[15] The inset shows a 2D reconstruction of the projection of the nanocrystal using the circled Bragg reflection. The Miller indices were determined from a corresponding high-angle pattern.
Reprinted by permission from Springer Customer Service Center GmbH: Nature 470, Femtosecond X-ray protein nanocrystallography, Chapman, H. N. *et al.* 2011.

Hattne *et al.* (2014) improved their SFX indexing yield by providing a target unit cell from the known crystal structure, and implementing indexing of multiple crystals per shot.[112] Nearby Bragg spots were used to empirically estimate 2D spot shapes (*i.e.* tight integration masks), and integrated intensities were scaled to an isomorphous reference structure.[113] The tight integration masks rendered the lattice model particularly sensitive to accurate metrology. The individual resolution cutoff (based on the pattern's Wilson plot) minimized the integration of noise. A three-parameter model for spot shape (modeling bandwidth and mosaicity) was also suggested as a potential approach to improving integration accuracy. The model Bragg reflection profile is different for each of the diffraction patterns, so model fitting, iteration and optimization must be repeated for each shot. However, significant improvement in SFX data quality was obtained even prior to implementing any classical crystallographic 'post-refinement', by refinement of the multipanel detector metrology (to sub-pixel accuracy), and by applying a resolution cut-off to each lattice, despite much lower multiplicities at high resolution.[113]

Using simulations, White (2014) demonstrated the potential of post-refinement for improving SFX data with partiality defined as the volume fraction of spherical reciprocal lattice points intersecting a nest of equally weighted Ewald spheres,[114] a model similar to Rossmann *et al.* (1979).[108] Fully integrated intensities were initially estimated from the merged, scaled dataset, in which reflection partialities were based on initial indexing results (*i.e.* no external intensities were used). Iterative refinement of each pattern's diffraction geometry was carried out until the corrected intensities in each pattern best matched the latest merged intensities. The efficacy of partiality corrections critically depends on accurate indexing. However, the crystal orientation is not uniquely determined by the Bragg spot positions in snapshot diffraction patterns, requiring orientation refinement of the crystals to recover accurate structure factors.[115] Various crystal orientation refinement approaches have thus now been implemented in all SFX processing programs.[100,116–118]

By including the 'Ewald offset correction' (minimum rotation required to bring the maximum of a Bragg reflection to a monochromatic Ewald sphere[118]), Sauter *et al.* (2014) improved upon classical crystal orientation refinement.[115] They then used analytical least squares curve-fitting for two parameters, η_{eff}, the represented variations in beam divergence, as well as rotation and unit cell differences in mosaic blocks, and D_{eff}, the effective crystal size normal to the beam. Both D_{eff} and η_{eff} broaden the Bragg spot (assumed to be spherical), and the area of intersection between the Bragg spot volume and the monochromatic Ewald sphere can thus be expressed as eqn (8.3)[107,116]

$$A = \pi r_p^2 = \pi \left[\left(\frac{1}{D_{\text{eff}}} + \frac{n_{\text{eff}}}{2d} \right)^2 - r_h^2 \right] \tag{8.3}$$

as shown in Figure 8.5. The 3D partiality is then defined as area A normalized against the Bragg spot volume.

Building on the work by Sauter *et al.* (2014),[115] Uervirojnangkoorn *et al.* (2015) implemented classical post refinement, specifically adapted for XFEL data, which significantly decreased the number of required patterns necessary.[107]

Kabsch (2014) defined an explicit description of diffraction geometry, beam divergence, detector parameters, crystal mosaicity and an isotropic B factor, specifically for snapshot diffraction patterns.[118] The presented post-refinement algorithm involves iterative refinement of these parameters and various weighting factors, and 2D profile fitting to estimate observed intensities.[118] A monochromatic beam and Gaussian rocking curve were assumed. This approach showed lower indexing rates (and resolution), comparable overall merged data quality, yet lower B factors and more readily automated model building when compared to data processed by Monte Carlo integration (with no resolution cutoff for each pattern).[119]

Taking account of the X-ray bandwidth, crystal size, orientation, and mosaicity also led to an effective refinement method that enabled structure

solution from cypovirus polyhedra protein nanocrystals,[117] which was not possible from synchrotron cryo-cooled crystal data. Ginn *et al.* (2015) refined auto-indexed crystal orientations by calculating the bandwidth required to bring delta-function Bragg spots onto the Ewald sphere, iteratively minimizing the effective bandwidth for each pattern, then fitting it to a Gaussian.[117] Orientation matrix refinement, partiality modeling, scaling and stringent, empirical rejection thresholds significantly reduced the multiplicity needed for high-quality XFEL data at high resolution. By accurately modeling experimental parameters and partiality models of increasing complexity they have demonstrated that the quality of SFX data is much higher than that accessible with Monte Carlo integration.[120]

Some crystal symmetries give rise to an indexing ambiguity in SFX, such as for crystals where twinning by merohedry is possible,[102] which was problematic in the early years of SFX. Indexing ambiguities no longer poses a problem as they can be resolved by algorithms based on intensity correlations,[121,122] variations of which have been implemented in *CrystFEL*[102] and nXDS.[118]

8.5.5 Phasing and Model Refinement

Once a merged reflection list is obtained, phasing and model refinement is then carried out with conventional crystallographic software. The minimum number of indexed patterns required for a full dataset is highly variable. For static structures that can be solved with molecular replacement, the required number of patterns has decreased by an order of magnitude since the first structures were published. Most of this improvement is due to numerous developments in data analysis and sub-pixel accuracy in detector geometry refinement. A rough estimate of the minimum number of patterns required for a full SFX dataset is 10 000, which can be collected from only 30 minutes at 120 Hz, if we assume a hit rate of 10% and indexing rate of 50%. Hit rates, diffraction quality and the minimum number of indexed patterns needed is very sample dependent.

During the first few years of SFX at LCLS, user groups always involved developers of at least one of these software packages. Software development was part of the experiment and users had to keep up with frequent detector format updates and complicated detector geometry. Sample delivery, source characteristics, data collection and analysis were all far from the user-friendly environment crystallographers may come to expect at established synchrotron MX beamlines, including a stable X-ray beam and no need to worry about detector geometry and gain artifacts. The rapid developments in SFX tools, due to years of effort by XFEL facilities and user groups alike, have enabled SFX experiments to now be carried out by much smaller teams. For risky experiments where sample quality is not guaranteed, LCLS offers short shifts as 'protein crystal screening' (PCS) beamtimes. This allows many more users to explore the usefulness of SFX for their protein targets. It is indeed possible to collect a full dataset from a single PCS beamtime (typically

6 hours) if the experiment runs smoothly. Good conditions include no loss of XFEL beam, no nozzle clogging, no detector damage (*e.g.* due to strong diffraction from salt crystals in the buffer), and, ultimately, high quality microcrystals at an adequate concentration for a high hit and indexing rate. It is also not uncommon that no useful diffraction is collected from a PCS beamtime.

8.5.6 *De novo* Phasing of SFX Data

Initial attempts at *de novo* phasing of SFX data were particularly difficult, owing to the inherent inaccuracy of Monte Carlo integration of SFX data. Only a handful of successful demonstrations of *de novo* phasing of SFX data have been published to date and are mentioned below. Large SFX datasets were necessary to obtain sufficient anomalous signal from averaged intensities of symmetry-related reflections due to averaging over stochastic beam spectrum and intensity fluctuations, sample anisomorphism and partially recorded reflections of randomly-oriented crystals. However, as testament to the rapid progress in the accuracy and efficiency of SFX data processing, demonstrations of *de novo* phasing of SFX data are appearing in the literature with increasing structure complexity and decreasing demands on data volumes.

First, single-wavelength anomalous diffraction was used to phase gadolinium-derivatized lysozyme as a model system[53] using 8.5 keV X-rays at LCLS. A correlation between number of patterns and peak heights in the Patterson map was observed, and SHELXD[123] was able to find the correct heavy atom sites from as few as 7500 patterns. However, despite the strong anomalous signal from gadolinium, a highly redundant 2.1 Å-resolution dataset based on ~60 000 microcrystals (<2 μm^3 in volume each), was necessary for sufficiently accurate merged structure factors for an automatically traceable electron density map.

Single isomorphous replacement with anomalous scattering (SIRAS) makes use of larger isomorphous differences ($\Delta|F| \gtrsim 10\%$) than single-wavelength anomalous dispersion (SAD) where $\Delta|F| \lesssim 10\%$, in addition to the anomalous signal.[124] Attempts to use only SAD on Hg-derivative data from a test soluble protein, a luciferin-regenerating enzyme (LRE), were unsuccessful even with high redundancy to 1.6 Å with more than 80 000 indexed patterns.[124] However, the use of SIRAS made structure solution possible with only ~10 000 patterns of native and Hg-derivative data (each) at 1.7 Å resolution.[124] This important result benefited from the high degree of isomorphism ($<0.2\%$ variation in unit cell parameters) between native and Hg-derivative enzyme crystals, which is atypical, and efficient incorporation of heavy atoms into protein crystals is often unattainable.

As the SFX technique matured, it became increasingly clear where major sources of uncertainty in large SFX datasets lie. Careful refinement of the metrology of the multi-tiled detector at CXI, LCLS (the 64-panel Cornell-SLAC Pixel Array Detector (CSPAD)[125] depicted in Figure 8.2), coupled with systematic, manual detector-sample distance refinement (minimizing the

width of the Gaussian-like unit cell distributions) as well as numerous updates in the analysis software and data scaling later enabled successful *de novo* phasing from these Gd-lysozyme data from only ~7000 indexed patterns,[126] compared to the originally required 60 000 patterns. Similarly, in more recent work, Yamashita *et al.* (2017) demonstrated that improvements in the accuracy of spot prediction and intensity scaling enabled the previously unsuccessful SAD phasing of LRE (using 1.5 Å resolution data from 11 000 indexed patterns).[127]

Selenium derivatization, replacing methionine residues with selenomethionine residues, is very commonly used at synchrotrons for *de novo* phasing using the Se K-edge. Se-SAD phasing from SFX data was first demonstrated on selenobiotinyl-streptavidin microcrystals measured at LCLS with 480 000 indexed patterns in the final dataset.[128] XFEL pulses spanning an order of magnitude in intensity (between 0.093 mJ and 0.93 mJ of 12.8 keV photons pulse^{-1}) were used to obtain a large dynamic range in the merged dataset. The weak measured anomalous signal (~1%), exacerbated in part by the increased scattering background of the viscous delivery medium, and the absence of alpha helices for auto-building, rendered solutions using automated pipelines ineffective,[128] and required at least 300 000 patterns for successful phasing. The final successful *de novo* phasing from these challenging data does indicate that Se-SAD can potentially play a big role in *de novo* phasing from SFX data in the future. Recently, improvements in intensity scaling in *CrystFEL*[102] (even without any partiality modeling) enabled Se-SAD phasing using much fewer indexed patterns (only 13 000 and 60 000 for 2 different samples) from SFX data collected at sub-1 Å X-ray wavelength at SACLA.[127]

Native SAD is particularly difficult as the anomalous signals of sulfur or phosphorus are on the order of ~1–2%. The earliest published native SAD SFX phasing demonstration relied on anomalous signal from intrinsic sulfur atoms and a bound chlorine in lysozyme, using 7 keV X-rays at SACLA.[129] This work required a large SFX dataset of about 150 000 indexed patterns to reach sufficient anomalous signal. Sulfur SAD (S-SAD) phasing was also demonstrated on another test case, the soluble protein thaumatin, with SFX data collected at LCLS using 6 keV X-rays LCLS.[126] In the thaumatin case, automatic model building was possible from at least 125 000 indexed patterns.

Nakane *et al.* then demonstrated iodine-SAD (using the magic triangle I3C), SIR and SIRAS on the membrane protein bacteriorhodopsin (bR).[130] Iodine-SAD was possible with data up to 2.1 Å resolution using only 3000 and 4000 indexed patterns from native and derivative microcrystals, respectively. Importantly for membrane proteins, which often diffract only to medium resolution, they demonstrated SIR/SIRAS phasing was possible from medium resolution (~3.3 Å) SFX data. Their success relied on successful binding *in crystallo* of a heavy-atom additive 13a (HAD13a) comprised of I3C and a hydrophobic alkyl chain (to enhance its membrane protein affinity).

Importantly, the first S-SAD phasing of SFX data from a GPCR was reported in 2016[131] from microcrystals of the human A_{2A} adenosine receptor.

The anomalous data were collected at LCLS using 6 keV X-rays to maximize anomalous signal without significant compromise in resolution (~2.1 Å), and higher resolution data were collected with 1.27 Å X-rays but truncated at 1.9 Å, based on correlation coefficient (CC*). Here, S-SAD phasing relied on at least 500 000 merged indexed patterns. SFX native SAD was also used for SFX structure determination of a copper-nitride reductase and was easier and much less sensitive to parameters used in SHELXD trials, requiring fewer than 100 SHELXD trials compared to S-SAD which required more than 320 000 trials.[132] However, copper and other heavy metals are not particularly common in macromolecules.

Regardless of the software, target sample or facility used, successful *de novo* phasing of SFX data has so far required large data volumes (high redundancy) to obtain the necessary accuracy in anomalous signal. With the development of various SFX data analysis software and algorithms that go beyond Monte Carlo integration (including crystal orientation refinement, intensity scaling and partiality analysis), much smaller data volumes are required for structure solution and the use of *de novo* phasing has been able to move beyond model systems to solving new structures. As one such example, microcrystals of a naturally occurring paracrystalline larvicide, whose tiny size (~50 unit cells on a side) prevented structure determination at synchrotron sources, were successfully phased using multiple isomorphous replacement with anomalous scattering (MIRAS) and, critically, post-refinement (*cctbx.xfel* for structure factor determination[92] and *cctbx.prime* for post-refinement[107]).[133]

A number of FELs have begun offering a new two-color XFEL pulse structure with two narrow-bandwidth pulses at different energies and controllable time delay. This new mode of operation has enabled interesting new approaches to SFX pump-probe and MAD experiments. At LCLS, for example, using twin electron bunches can provide high-intensity two-color XFEL pulses with separation of ~120 eV and up to a few % possible, with controllable separation of up to tens of fs.[134] The two-color mode at SACLA, on the other hand, uses a split undulator to achieve a range of energy separation of >30% with precisely controlled attosecond delays and <10 fs pulses.[135] Recently, this new mode of operation at SACLA was used for SAD and MAD SFX phasing of the model system Gd-lysozyme again, but now requiring only ~5000 indexed patterns for automated model building.[136] Importantly, large energy separation leads to well-separated diffraction spots, so no deconvolution of partially-recorded intensities is necessary, and the tunability allows optimization to absorption edges of interest.

8.5.7 SFX Data Volumes and Data Sharing

The Coherent X-ray Imaging Data Bank (CXIDB, www.cxidb.org)[137] was established for sharing of datasets with the broader scientific community, along with scripts and necessary metadata to enable reproduction of published analysis of the uploaded data. Multiple XFEL datasets, from SFX and

single particle experiments, have been uploaded to CXIDB. An open access online journal, Scientific Data (Nature Publishing) was launched in May 2014, for sharing of datasets for the reuse of data.[138–141] For the young field of SFX, this is a very valuable addition to CXIDB, that facilitates the sharing of data, metadata and scripts necessary to reproduce and validate published analysis and enables research groups worldwide to test and develop algorithms.

8.6 New Developments

8.6.1 Sparse Crystal Pattern Indexing

Autoindexing typically fails for patterns with fewer than 15 Bragg reflections, meaning weak diffraction patterns (*e.g.* from the smallest crystals or small crystal volume intersected by XFEL pulse) are wasted. For novel structures, integrating reflections from every crystal hit, even low-resolution patterns, could be crucial for structure solution. Algorithms are being developed specifically for sparse patterns based on the EMC (expansion, maximization, compression) algorithm developed for orientation determination in single particle diffraction imaging.[97,142]

8.6.2 Nanocrystal Shape Transform Phasing

Novel direct phasing methods for crystals are being developed making use of the high spatial coherence of XFEL pulses. The goal of 'shape transform phasing' is to solve the nanocrystal phase problem *ab initio* by sampling intensities at finer intervals than the conventional Bragg condition allows.[143,144]

Simulations have confirmed that the suggested method of dividing-out the averaged shape transform from diffraction patterns sampled at spacings finer than the Bragg reflections[143] is able to achieve *ab initio* phase reconstruction of the unit cell.[145,146] Experimental tests have confirmed this approach at medium resolution using soft X-rays and 2D artificial crystals formed by lithography as shown in Figure 8.7A.[144] The shape transform approach could thus provide an extremely valuable model-independent phasing method that does not require atomic-resolution data or chemical modifications to the sample.

8.6.3 Continuous Diffuse Scattering

One of the most interesting recent realizations was that low/medium resolution Bragg diffraction does not necessarily reflect heterogeneity between the proteins. The loss of crystalline order by random translations of rigid bodies (which could be the asymmetric unit or a multimer) away from ideal lattice positions will lead to X-rays scattered away from Bragg spots, while the integrated scattered intensity remains constant.[29,147] Provided the rigid

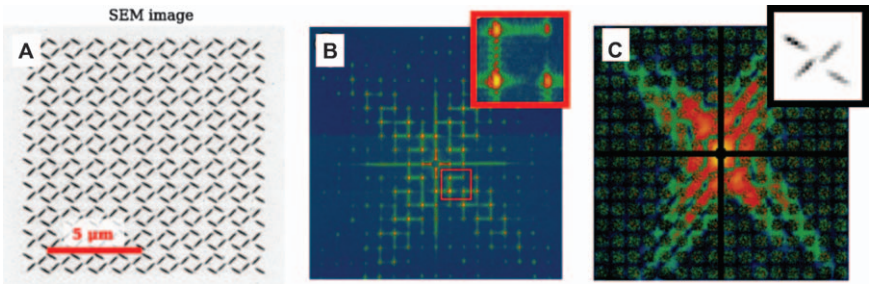


Figure 8.7 (A) Example of a fabricated 2D crystal used in the first experimental verification of phase retrieval using shape transforms. (B) Diffraction pattern observed for the crystal in (A), with the red box showing an enlarged region containing visible scattering from its shape transform. (C) A demodulated pattern showing the molecular transform and its corresponding reconstruction on the upper right.
Figure adapted from Kirian *et al.* Direct Phasing of Finite Crystals Illuminated with a Free-Electron Laser, *Phys. Rev. X*, **5**, 2015, licensed under CC-BY 3.0.

bodies are identical and without significant rotational disorder, the continuous diffuse diffraction pattern that emerges from between the Bragg reflections is that of the continuous Fourier transform of the rigid bodies. This observation was used to extend the crystal resolution from photosystem II microcrystals from 4.5 Å from Bragg data alone, to 3.5 Å using iterative phasing of the 'oversampled' diffuse scattering beyond the Bragg spots.[147–149] Efforts are underway to determine the degree of generality of this technique for *ab initio* phasing of other membrane protein crystals that show the same type of continuous diffuse scattering.[149]

8.6.4 Single-layer 2D Crystals

Many membrane proteins naturally form 2D crystals, as studied extensively by cryo-electron microscopy. The first 2D X-ray crystallography patterns were obtained by using an XFEL, since insufficient flux and radiation damage limits prevent this when using synchrotron radiation. These XFEL patterns, from a single-layer two-dimensional membrane protein crystal (bacteriorhodopsin mutant) showed diffraction to 8.5 Å,[63] which was extended to 7 Å by accurate merging[62] and, with further analysis, it was shown that reliable intensities of reflections at 4 Å were obtainable after about ~200 observations (from single-layer 2D crystals). This demonstration of 4 Å resolution from a single layer of a membrane protein crystal, at room temperature, shows that static or time-resolved studies could soon be possible with XFELs on 2D crystals, probing biochemically important (functionally relevant) large-scale motions that may be suppressed in 3D crystals.[64] If a sufficient number of patterns can be obtained from 2D crystals, it may be possible to phase the data by using a compact support constraint normal to the crystalline monolayer (imposing zero electron density on either side of

the 2D crystal).[150] Membrane protein crystals with high solvent content (>50%) may be amenable to iterative phasing from the Bragg data alone.[151]

8.6.5 Incoherent Diffractive Imaging

Recently, Classen *et al.* (2017) presented a new incoherent diffractive imaging (IDI) technique that uses intensity correlations of incoherently scattered photons to determine the 3D atomic structure from crystals or single particles.[152] Simulations based on fluorescence emission show that provided the X-ray pulses are shorter than the coherence time of *K*-shell fluorescence photons (*i.e.* their radiative lifetimes which are in the order of a few femtoseconds), the scattered X-rays in each shot have a fixed phase relation with the incident X-rays. Thus, the 3D atomic-resolution structure of the fluorescing atoms can be derived, analogous to substructure determination in anomalous phasing. For the same experimental geometry, IDI provides twice the resolution of crystallography or coherent diffractive imaging experiments, and because fluorescence is often the dominant interaction, flux requirements are lower. The requirements of high brilliance and ultrashort pulses and high repetition rates are already met by existing XFELs. The realization of the proposed technique would open up a fundamentally new, superior approach to atomic imaging, that would benefit crystallography and single particle imaging.[152]

8.7 Conclusion

The current goals in SFX include improvements in sample delivery to further optimize hit rates, detector calibrations, obtaining reliable spectrum measurements for each shot, developing novel phasing methods (finite crystals, continuous diffuse scattering, 2D crystal phasing to name a few), sparse diffraction indexing, improving integration of very finely spaced spots from particularly large membrane protein unit cells and improving mix-and-inject injectors to improve the time resolution. Concurrently, the last few years have seen data analysis and sample delivery methods developed for serial crystallography at XFELs being increasingly adopted at synchrotron sources.[45,67,153–155] High-intensity microfocus synchrotron beamlines are being used for structure determination of ever-smaller crystals, with automated microcrystal selection and data collection from multiple microcrystals on a pin, for example, and liquid/LCP jet and fixed-target serial millisecond crystallography at synchrotrons will continue to evolve.[67,153,156]

XFEL crystallography has faced challenges and challengers, but SFX persisted and thrived. Though the technique is younger than a decade, there is no question that SFX has enabled and demonstrated unprecedented opportunities through room-temperature, atomic-resolution, time-resolved femtosecond-scale pump-probe or millisecond mix-and-inject crystallography. Static structure determination from microcrystals of soluble proteins is no longer at the forefront of SFX development. Many in the

rapidly-growing SFX community are optimizing methods for studying macromolecule dynamics through time-resolved SFX (optionally with simultaneous spectroscopy), improving the time resolution in mix-and-inject SFX, establishing robust *de novo* phasing methods, increasingly understanding, mitigating (or exploiting) radiation damage, and exploring novel techniques making use of the unique properties of XFEL beams. XFEL beamtime for structural biology will continue to be very exciting and competitive, as unprecedented experimental modalities are invented and it is inspiring to consider what great insights in structural biology, enzymology and drug development the next decade of SFX will bring.

Acknowledgements

The support from US National Science Foundation (NSF) Science and Technology Center 'BioXFEL' Award STC-1231306 and the NSF Award #1565180 ABI Innovation: New Algorithms for Biological X-ray Free Electron Laser Data are acknowledged.

References

1. C. Bostedt, S. Boutet, D. M. Fritz, Z. Huang, H. J. Lee, H. T. Lemke, A. Robert, W. F. Schlotter, J. J. Turner and G. J. Williams, *Rev. Mod. Phys.*, 2016, **88**, 207.
2. I. Schlichting, *IUCrJ*, 2015, **2**, 246–255.
3. J. C. H. Spence, *IUCrJ*, 2017, **4**, 322–339.
4. C. Pellegrini, *Eur. Phys. J. H*, 2012, **37**, 659–708.
5. S. Boutet, L. Foucar, T. R. M. Barends, S. Botha, R. B. Doak, J. E. Koglin, M. M. Messerschmidt, K. Nass, I. Schlichting, M. M. Seibert, R. L. Shoeman and G. J. Williams, *J. Synchrotron Radiat.*, 2015, **22**, 1–10.
6. M. R. Howells, T. Beetz, H. N. Chapman, C. Cui, J. M. Holton, C. J. Jacobsen, J. Kirz, E. Lima, S. Marchesini, H. Miao, D. Sayre, D. A. Shapiro, J. C. H. Spence and D. Starodub, *J. Electron Spectrosc. Relat. Phenom.*, 2009, **170**, 4–12.
7. R. Neutze, R. Wouts, D. van der Spoel, E. Weckert and J. Hajdu, *Nature*, 2000, **406**, 752–757.
8. R. Neutze, *Phil. Trans. R. Soc., B*, 2014, **369**, 20130318.
9. A. Barty, C. Caleman, A. Aquila, N. Timneanu, L. Lomb, T. A. White, J. Andreasson, D. Arnlund, S. Bajt, T. R. M. Barends, M. Barthelmess, M. J. Bogan, C. Bostedt, J. D. Bozek, R. Coffee, N. Coppola, J. Davidsson, D. P. DePonte, R. B. Doak, T. Ekeberg, V. Elser, S. W. Epp, B. Erk, H. Fleckenstein, L. Foucar, P. Fromme, H. Graafsma, L. Gumprecht, J. Hajdu, C. Y. Hampton, R. Hartmann, A. Hartmann, G. Hauser, H. Hirsemann, P. Holl, M. S. Hunter, L. Johansson, S. Kassemeyer, N. Kimmel, R. A. Kirian, M. Liang, F. R. N. C. Maia, E. Malmerberg, S. Marchesini, A. V. Martin, K. Nass, R. Neutze, C. Reich, D. Rolles, B. Rudek, A. Rudenko, H. Scott, I. Schlichting, J. Schulz, M. M. Seibert,

R. L. Shoeman, R. G. Sierra, H. Soltau, J. C. H. Spence, F. Stellato, S. Stern, L. Strüder, J. Ullrich, X. Wang, G. Weidenspointner, U. Weierstall, C. B. Wunderer and H. N. Chapman, *Nat. Photonics*, 2011, **6**, 35–40.
10. H. N. Chapman, C. Caleman and N. Timneanu, *Phil. Trans. R. Soc., B*, 2014, **369**, 20130313.
11. A. Barty, J. Küpper and H. N. Chapman, *Annu. Rev. Phys. Chem.*, 2013, **64**, 415–435.
12. C. Caleman, G. Huldt, F. R. N. C. Maia, C. Ortiz, F. G. Parak, J. Hajdu, D. van der Spoel, H. N. Chapman and N. Timneanu, *ACS Nano*, 2011, **5**, 139–146.
13. R. L. Owen, E. Rudino-Pinera and E. F. Garman, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 4912–4917.
14. J. Yano, J. Kern, K.-D. Irrgang, M. J. Latimer, U. Bergmann, P. Glatzel, Y. Pushkar, J. Biesiadka, B. Loll, K. Sauer, J. Messinger, A. Zouni and V. K. Yachandra, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 12047–12052.
15. H. N. Chapman, P. Fromme, A. Barty, T. A. White, R. A. Kirian, A. Aquila, M. S. Hunter, J. Schulz, D. P. DePonte, U. Weierstall, R. B. Doak, F. R. N. C. Maia, A. V. Martin, I. Schlichting, L. Lomb, N. Coppola, R. L. Shoeman, S. W. Epp, R. Hartmann, D. Rolles, A. Rudenko, L. Foucar, N. Kimmel, G. Weidenspointner, P. Holl, M. Liang, M. Barthelmess, C. Caleman, S. Boutet, M. J. Bogan, J. Krzywinski, C. Bostedt, S. Bajt, L. Gumprecht, B. Rudek, B. Erk, C. Schmidt, A. Hömke, C. Reich, D. Pietschner, L. Strüder, G. Hauser, H. Gorke, J. Ullrich, S. Herrmann, G. Schaller, F. Schopper, H. Soltau, K.-U. Kühnel, M. M. Messerschmidt, J. D. Bozek, S. P. Hau-Riege, M. Frank, C. Y. Hampton, R. G. Sierra, D. Starodub, G. J. Williams, J. Hajdu, N. Timneanu, M. M. Seibert, J. Andreasson, A. Rocker, O. Jönsson, M. Svenda, S. Stern, K. Nass, R. Andritschke, C.-D. Schröter, F. Krasniqi, M. Bott, K. E. Schmidt, X. Wang, I. Grotjohann, J. M. Holton, T. R. M. Barends, R. Neutze, S. Marchesini, R. Fromme, S. Schorb, D. Rupp, M. Adolph, T. Gorkhover, I. Andersson, H. Hirsemann, G. Potdevin, H. Graafsma, B. Nilsson and J. C. H. Spence, *Nature*, 2011, **469**, 73–77.
16. C. Gati, D. Oberthuer, O. M. Yefanov, R. D. Bunker, F. Stellato, E. Chiu, S.-M. Yeh, A. Aquila, S. Basu, R. Bean, K. R. Beyerlein, S. Botha, S. Boutet, D. P. DePonte, R. B. Doak, R. Fromme, L. Galli, I. Grotjohann, D. R. James, C. Kupitz, L. Lomb, M. M. Messerschmidt, K. Nass, K. Rendek, R. L. Shoeman, D. Wang, U. Weierstall, T. A. White, G. J. Williams, N. A. Zatsepin, P. Fromme, J. C. H. Spence, K. N. Goldie, J. A. Jehle, P. Metcalf, A. Barty and H. N. Chapman, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, 2247–2252.
17. S.-K. Son, L. Young and R. Santra, *Phys. Rev. A*, 2011, **83**, 033402.
18. K. Nass, L. Foucar, T. R. M. Barends, E. Hartmann, S. Botha, R. L. Shoeman, R. B. Doak, R. Alonso-Mori, A. Aquila, S. Bajt, A. Barty, R. Bean, K. R. Beyerlein, M. Bublitz, N. Drachmann, J. Gregersen, H. O. Jonsson, W. Kabsch, S. Kassemeyer, J. E. Koglin, M. Krumrey,

D. Mattle, M. M. Messerschmidt, P. Nissen, L. Reinhard, O. Sitsel, D. Sokaras, G. J. Williams, S. Hau-Riege, N. Timneanu, C. Caleman, H. N. Chapman, S. Boutet and I. Schlichting, *J. Synchrotron Radiat.*, 2015, **22**, 1–14.

19. J. Kern, R. Alonso-Mori, R. Tran, J. Hattne, R. J. Gildea, N. Echols, C. Glöckner, J. Hellmich, H. Laksmono and R. G. Sierra, *Science*, 2013, **340**, 491–495.
20. J. Kern, V. K. Yachandra and J. Yano, *Curr. Opin. Struct. Biol.*, 2015, **34**, 87–98.
21. F. D. Fuller, S. Gul, R. Chatterjee, E. S. Burgie, I. D. Young, H. Lebrette, V. Srinivas, A. S. Brewster, T. Michels-Clark, J. A. Clinger, B. Andi, M. Ibrahim, E. Pastor, C. de Lichtenberg, R. Hussein, C. J. Pollock, M. Zhang, C. A. Stan, T. Kroll, T. Fransson, C. Weninger, M. Kubin, P. Aller, L. Lassalle, P. Bräuer, M. D. Miller, M. Amin, S. Koroidov, C. G. Roessler, M. Allaire, R. G. Sierra, P. T. Docker, J. M. Glownia, S. Nelson, J. E. Koglin, D. Zhu, M. Chollet, S. Song, H. Lemke, M. Liang, D. Sokaras, R. Alonso, Mori, A. Zouni, J. Messinger, U. Bergmann, A. K. Boal, J. M. Bollinger, C. Krebs, M. Högbom, G. N. Phillips, R. D. Vierstra, N. K. Sauter, A. M. Orville, J. Kern, V. K. Yachandra and J. Yano, *Nat. Methods*, 2017, **14**, 443–449.
22. J. S. Fraser, H. van den Bedem, A. J. Samelson, P. T. Lang, J. M. Holton, N. Echols and T. Alber, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 16247–16252.
23. L. Galli, S. K. Son, T. A. White, R. Santra, H. N. Chapman and M. H. Nanao, *J. Synchrotron Radiat.*, 2015, **22**, 1–7.
24. H. N. Chapman, C. Caleman and N. Timneanu, *Phil. Trans. R. Soc., B*, 2014, **369**, 20130313.
25. R. Neutze, G. Brändén and G. F. Schertler, *Curr. Opin. Struct. Biol.*, 2015, **33**, 115–125.
26. J. Kern, R. Alonso-Mori, J. Hellmich, R. Tran, J. Hattne, H. Laksmono, C. Glockner, N. Echols, R. G. Sierra, J. Sellberg, B. Lassalle-Kaiser, R. J. Gildea, P. Glatzel, R. W. Grosse-Kunstleve, M. J. Latimer, T. A. McQueen, D. DiFiore, A. R. Fry, M. M. Messerschmidt, A. Miahnahri, D. W. Schafer, M. M. Seibert, D. Sokaras, T. C. Weng, P. H. Zwart, W. E. White, P. D. Adams, M. J. Bogan, S. Boutet, G. J. Williams, J. Messinger, N. K. Sauter, A. Zouni, U. Bergmann, J. Yano and V. K. Yachandra, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 9721–9726.
27. J. Kern, R. Tran, R. Alonso-Mori, S. Koroidov, N. Echols, J. Hattne, M. Ibrahim, S. Gul, H. Laksmono, R. G. Sierra, R. J. Gildea, G. Han, J. Hellmich, B. Lassalle-Kaiser, R. Chatterjee, A. S. Brewster, C. A. Stan, C. G. O. ckner, A. Lampe, D. O. R. DiFiore, D. Milathianaki, A. R. Fry, M. M. Seibert, J. E. Koglin, E. Gallo, J. Uhlig, D. Sokaras, T.-C. Weng, P. H. Zwart, D. E. Skinner, M. J. Bogan, M. M. Messerschmidt, P. Glatzel, G. J. Williams, S. Boutet, P. D. Adams, A. Zouni, J. Messinger, N. K. Sauter, U. Bergmann, J. Yano and V. K. Yachandra, *Nat. Commun.*, 2014, **5**, 1–11.

28. C. Kupitz, S. Basu, I. Grotjohann, R. Fromme, N. A. Zatsepin, K. N. Rendek, M. S. Hunter, R. L. Shoeman, T. A. White, D. Wang, D. James, J.-H. Yang, D. E. Cobb, B. Reeder, R. G. Sierra, H. Liu, A. Barty, A. L. Aquila, D. Deponte, R. A. Kirian, S. Bari, J. J. Bergkamp, K. R. Beyerlein, M. J. Bogan, C. Caleman, T.-C. Chao, C. E. Conrad, K. M. Davis, H. Fleckenstein, L. Galli, S. P. Hau-Riege, S. Kassemeyer, H. Laksmono, M. Liang, L. Lomb, S. Marchesini, A. V. Martin, M. M. Messerschmidt, D. Milathianaki, K. Nass, A. Ros, S. Roy-Chowdhury, K. Schmidt, M. Seibert, J. Steinbrener, F. Stellato, L. Yan, C. H. Yoon, T. A. Moore, A. L. Moore, Y. Pushkar, G. J. Williams, S. Boutet, R. B. Doak, U. Weierstall, M. Frank, H. N. Chapman, J. C. H. Spence and P. Fromme, *Nature*, 2014, **513**, 261–265.
29. K. Ayyer, O. M. Yefanov, D. Oberthür, S. Roy-Chowdhury, L. Galli, V. Mariani, S. Basu, J. Coe, C. E. Conrad, R. Fromme, A. Schaffer, K. Dorner, D. James, C. Kupitz, M. Metz, G. Nelson, P. L. Xavier, K. R. Beyerlein, M. Schmidt, I. Sarrou, J. C. H. Spence, U. Weierstall, T. A. White, J.-H. Yang, Y. Zhao, M. Liang, A. Aquila, M. S. Hunter, J. S. Robinson, J. E. Koglin, S. Boutet, P. Fromme, A. Barty and H. N. Chapman, *Nature*, 2016, **530**, 202–206.
30. I. Ishigami, N. A. Zatsepin, M. Hikita, C. E. Conrad, G. Nelson, J. D. Coe, S. Basu, T. D. Grant, M. H. Seaberg, R. G. Sierra, M. S. Hunter, P. Fromme, R. Fromme, S.-R. Yeh and D. L. Rousseau, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, 8011–8016.
31. K. Hirata, K. Shinzawa-Itoh, N. Yano, S. Takemura, K. Kato, M. Hatanaka, K. Muramoto, T. Kawahara, T. Tsukihara, E. Yamashita, K. Tono, G. Ueno, T. Hikima, H. Murakami, Y. Inubushi, M. Yabashi, T. Ishikawa, M. Yamamoto, T. Ogura, H. Sugimoto, J.-R. Shen, S. Yoshikawa and H. Ago, *Nat. Methods*, 2014, **11**, 734–736.
32. L. C. Johansson, D. Arnlund, G. Katona, T. A. White, A. Barty, D. P. DePonte, R. L. Shoeman, C. Wickstrand, A. Sharma, G. J. Williams, A. Aquila, M. J. Bogan, C. Caleman, J. Davidsson, R. B. Doak, M. Frank, R. Fromme, L. Galli, I. Grotjohann, M. S. Hunter, S. Kassemeyer, R. A. Kirian, C. Kupitz, M. Liang, L. Lomb, E. Malmerberg, A. V. Martin, M. M. Messerschmidt, K. Nass, L. Redecke, M. M. Seibert, J. S. O. hamn, J. Steinbrener, F. Stellato, D. Wang, W. Y. Wahlgren, U. Weierstall, S. Westenhoff, N. A. Zatsepin, S. Boutet, J. C. H. Spence, I. Schlichting, H. N. Chapman, P. Fromme and R. Neutze, *Nat. Commun.*, 2013, **4**, 1–7.
33. J. Xiang, E. Chun, C. Liu, L. Jing, Z. Al-Sahouri, L. Zhu and W. Liu, *Trends Pharmacol. Sci.*, 2016, **37**, 1055–1069.
34. M. Caffrey, *Acta Crystallogr. Sect. F*, 2015, **71**, 3–18.
35. S. G. F. Rasmussen, H.-J. Choi, J. J. Fung, E. Pardon, P. Casarosa, P. S. Chae, B. T. DeVree, D. M. Rosenbaum, F. S. Thian, T. S. Kobilka, A. Schnapp, I. Konetzki, R. K. Sunahara, S. H. Gellman, A. Pautsch, J. Steyaert, W. I. Weis and B. K. Kobilka, *Nature*, 2010, **469**, 175–180.

36. H. E. Kato, F. Zhang, O. Yizhar, C. Ramakrishnan, T. Nishizawa, K. Hirata, J. Ito, Y. Aita, T. Tsukazaki, S. Hayashi, P. Hegemann, A. D. Maturana, R. Ishitani, K. Deisseroth and O. Nureki, *Nature*, 2012, **482**, 369–374.
37. M. Caffrey, D. Li and A. Dukkipati, *Biochemistry*, 2012, **51**, 6266–6288.
38. U. Weierstall, D. James, C. Wang, T. A. White, D. Wang, W. Liu, J. C. H. Spence, R. B. Doak, G. Nelson, P. Fromme, R. Fromme, I. Grotjohann, C. Kupitz, N. A. Zatsepin, H. Liu, S. Basu, D. Wacker, G. W. Han, V. Katritch, S. Boutet, M. M. Messerschmidt, G. J. Williams, J. E. Koglin, M. M. Seibert, M. Klinker, C. Gati, R. L. Shoeman, A. Barty, H. N. Chapman, R. A. Kirian, K. R. Beyerlein, R. C. Stevens, D. Li, S. T. A. Shah, N. Howe, M. Caffrey and V. Cherezov, *Nat. Commun.*, 2014, **5**, 1–6.
39. W. Liu, D. Wacker, C. Gati, G. W. Han, D. James, D. Wang, G. Nelson, U. Weierstall, V. Katritch, A. Barty, N. A. Zatsepin, D. Li, M. M. Messerschmidt, S. Boutet, G. J. Williams, J. E. Koglin, M. M. Seibert, C. Wang, S. T. A. Shah, S. Basu, R. Fromme, C. Kupitz, K. N. Rendek, I. Grotjohann, P. Fromme, R. A. Kirian, K. R. Beyerlein, T. A. White, H. N. Chapman, M. Caffrey, J. C. H. Spence, R. C. Stevens and V. Cherezov, *Science*, 2013, **342**, 1521–1524.
40. D. Li, P. J. Stansfeld, M. S. P. Sansom, A. Keogh, L. Vogeley, N. Howe, J. A. Lyons, D. Aragao, P. Fromme, R. Fromme, S. Basu, I. Grotjohann, C. Kupitz, K. Rendek, U. Weierstall, N. A. Zatsepin, V. Cherezov, W. Liu, S. Bandaru, N. J. English, C. Gati, A. Barty, O. M. Yefanov, H. N. Chapman, K. Diederichs, M. M. Messerschmidt, S. Boutet, G. J. Williams, M. M. Seibert and M. Caffrey, *Nat. Commun.*, 2015, **6**, 1–12.
41. Y. Kang, X. E. Zhou, X. Gao, Y. He, W. Liu, A. Ishchenko, A. Barty, T. A. White, O. M. Yefanov, G. W. Han, Q. Xu, P. W. de Waal, J. Ke, M. H. E. Tan, C. Zhang, A. Moeller, G. M. West, B. D. Pascal, N. van Eps, L. N. Caro, S. A. Vishnivetskiy, R. J. Lee, K. M. Suino-Powell, X. Gu, K. Pal, J. Ma, X. Zhi, S. Boutet, G. J. Williams, M. M. Messerschmidt, C. Gati, N. A. Zatsepin, D. Wang, D. James, S. Basu, S. Roy-Chowdhury, C. E. Conrad, J. Coe, H. Liu, S. Lisova, C. Kupitz, I. Grotjohann, R. Fromme, Y. Jiang, M. Tan, H. Yang, J. Li, M. Wang, Z. Zheng, D. Li, N. Howe, Y. Zhao, J. Standfuss, K. Diederichs, Y. Dong, C. S. Potter, B. Carragher, M. Caffrey, H. Jiang, H. N. Chapman, J. C. H. Spence, P. Fromme, U. Weierstall, O. P. Ernst, V. Katritch, V. V. Gurevich, P. R. Griffin, W. L. Hubbell, R. C. Stevens, V. Cherezov, K. Melcher and H. E. Xu, *Nature*, 2015, **523**, 561–567.
42. H. Zhang, H. Unal, C. Gati, G. W. Han, W. Liu, N. A. Zatsepin, D. James, D. Wang, G. Nelson, U. Weierstall, M. R. Sawaya, Q. Xu, M. M. Messerschmidt, G. J. Williams, S. Boutet, O. M. Yefanov, T. A. White, C. Wang, A. Ishchenko, K. C. Tirupula, R. Desnoyer, J. Coe, C. E. Conrad, P. Fromme, R. C. Stevens, V. Katritch, S. S. Karnik and V. Cherezov, *Cell*, 2015, **161**, 833–844.

43. H. Zhang, A. Qiao, D. Yang, L. Yang, A. Dai, C. de Graaf, S. Reedtz-Runge, V. Dharmarajan, H. Zhang, G. W. Han, T. D. Grant, R. G. Sierra, U. Weierstall, G. Nelson, W. Liu, Y. Wu, L. Ma, X. Cai, G. Lin, X. Wu, Z. Geng, Y. Dong, G. Song, P. R. Griffin, J. Lau, V. Cherezov, H. Yang, M. A. Hanson, R. C. Stevens, Q. Zhao, H. Jiang, M.-W. Wang and B. Wu, *Nature*, 2017, **546**, 259–264.
44. G. Fenalti, N. A. Zatsepin, C. Betti, P. Giguere, G. W. Han, A. Ishchenko, W. Liu, K. Guillemyn, H. Zhang, D. James, D. Wang, U. Weierstall, J. C. H. Spence, S. Boutet, M. M. Messerschmidt, G. J. Williams, C. Gati, O. M. Yefanov, T. A. White, D. Oberthuer, M. Metz, C. H. Yoon, A. Barty, H. N. Chapman, S. Basu, J. Coe, C. E. Conrad, R. Fromme, P. Fromme, D. Tourwé, P. W. Schiller, B. L. Roth, S. Ballet, V. Katritch, R. C. Stevens and V. Cherezov, *Nat. Struct. Mol. Biol.*, 2015, **22**, 265–268.
45. C. Gati, G. Bourenkov, M. Klinge, D. Rehders, F. Stellato, D. Oberthur, O. M. Yefanov, B. P. Sommer, S. Mogk, M. Duszenko, C. Betzel, T. R. Schneider, H. N. Chapman and L. Redecke, *IUCrJ*, 2014, **1**, 1–8.
46. F. X. Gallat, N. Matsugaki, N. P. Coussens, K. J. Yagi, M. Boudes, T. Higashi, D. Tsuji, Y. Tatano, M. Suzuki, E. Mizohata, K. Tono, Y. Joti, T. Kameshima, J. Park, C. Song, T. Hatsui, M. Yabashi, E. Nango, K. Itoh, F. Coulibaly, S. Tobe, S. Ramaswamy, B. Stay, S. Iwata and L. M. G. Chavas, *Phil. Trans. R. Soc., B*, 2014, **369**, 20130497.
47. U. Weierstall, R. B. Doak, J. C. H. Spence, D. Starodub, D. Shapiro, P. Kennedy, J. Warner, G. G. Hembree, P. Fromme and H. N. Chapman, *Exp. Fluids*, 2007, **44**, 675–689.
48. U. Weierstall, J. C. H. Spence and R. B. Doak, *Rev. Sci. Instrum.*, 2012, **83**, 035108.
49. J. C. H. Spence and R. B. Doak, *Phys. Rev. Lett.*, 2004, **92**, 198102.
50. M. S. Hunter, B. Segelke, M. M. Messerschmidt, G. J. Williams, N. A. Zatsepin, A. Barty, W. H. Benner, D. B. Carlson, M. Coleman, A. T. Graf, S. P. Hau-Riege, T. Pardini, M. M. Seibert, J. Evans, S. Boutet and M. Frank, *Sci. Rep.*, 2014, **4**, 6026.
51. S. Boutet, L. Lomb, G. J. Williams, T. R. M. Barends, A. Aquila, R. B. Doak, U. Weierstall, D. P. DePonte, J. Steinbrener, R. L. Shoeman, M. M. Messerschmidt, A. Barty, T. A. White, S. Kassemeyer, R. A. Kirian, M. M. Seibert, P. A. Montanez, C. Kenney, R. Herbst, P. Hart, J. Pines, G. Haller, S. M. Gruner, H. T. Philipp, M. W. Tate, M. Hromalik, L. J. Koerner, N. van Bakel, J. Morse, W. Ghonsalves, D. Arnlund, M. J. Bogan, C. Caleman, R. Fromme, C. Y. Hampton, M. S. Hunter, L. C. Johansson, G. Katona, C. Kupitz, M. Liang, A. V. Martin, K. Nass, L. Redecke, F. Stellato, N. Timneanu, D. Wang, N. A. Zatsepin, D. Schafer, J. Defever, R. Neutze, P. Fromme, J. C. H. Spence, H. N. Chapman and I. Schlichting, *Science*, 2012, **337**, 362–364.
52. L. Redecke, K. Nass, D. P. DePonte, T. A. White, D. Rehders, A. Barty, F. Stellato, M. Liang, T. R. M. Barends, S. Boutet, G. J. Williams, M. M. Messerschmidt, M. M. Seibert, A. Aquila, D. Arnlund, S. Bajt, T. Barth, M. J. Bogan, C. Caleman, T.-C. Chao, R. B. Doak,

H. Fleckenstein, M. Frank, R. Fromme, L. Galli, I. Grotjohann, M. S. Hunter, L. C. Johansson, S. Kassemeyer, G. Katona, R. A. Kirian, R. Koopmann, C. Kupitz, L. Lomb, A. V. Martin, S. Mogk, R. Neutze, R. L. Shoeman, J. Steinbrener, N. Timneanu, D. Wang, U. Weierstall, N. A. Zatsepin, J. C. H. Spence, P. Fromme, I. Schlichting, M. Duszenko, C. Betzel and H. N. Chapman, *Science*, 2013, **339**, 227–230.

53. T. R. M. Barends, L. Foucar, S. Botha, R. B. Doak, R. L. Shoeman, K. Nass, J. E. Koglin, G. J. Williams, S. Boutet, M. M. Messerschmidt and I. Schlichting, *Nature*, 2014, **505**, 244–247.
54. A. Y. Lyubimov, T. D. Murray, A. Koehl, I. E. Araci, M. Uervirojnangkoorn, O. B. Zeldin, A. E. Cohen, S. M. Soltis, E. L. Baxter, A. S. Brewster, N. K. Sauter, A. T. Brunger and J. M. Berger, *Acta Crystallogr., Sect. D*, 2015, **71**, 928–940.
55. P. Roedig, I. Vartiainen, R. Duman, S. Panneerselvam, N. Stübe, O. Lorbeer, M. Warmer, G. Sutton, D. I. Stuart, E. Weckert, D. J. Vine, A. Wagner and A. Meents, *Sci. Rep.*, 2015, **5**, 10451.
56. E. L. Baxter, L. Aguila, R. Alonso-Mori, C. O. Barnes, C. A. Bonagura, W. Brehmer, A. T. Brunger, G. Calero, T. T. Caradoc-Davies, R. Chatterjee, W. F. Degrado, J. S. Fraser, M. Ibrahim, J. Kern, B. K. Kobilka, A. C. Kruse, K. M. Larsson, H. T. Lemke, A. Y. Lyubimov, A. Manglik, S. E. McPhillips, E. Norgren, S. S. Pang, S. M. Soltis, J. Song, J. Thomaston, Y. Tsai, W. I. weis, R. A. Woldeyes, V. Yachandra, J. Yano, A. Zouni and A. E. Cohen, *Acta Crystallogr., Sect. D*, 2016, **72**, 2–11.
57. P. Roedig, H. M. Ginn, T. Pakendorf, G. Sutton, K. Harlos, T. S. Walter, J. Meyer, P. Fischer, R. Duman, I. Vartiainen, B. Reime, M. Warmer, A. S. Brewster, I. D. Young, T. Michels-Clark, N. K. Sauter, A. Kotecha, J. Kelly, D. J. Rowlands, M. Sikorsky, S. Nelson, D. S. Damiani, R. Alonso-Mori, J. Ren, E. E. Fry, C. David, D. I. Stuart, A. Wagner and A. Meents, *Nat. Methods*, 2017, **14**, 805–810.
58. C. Mueller, A. Marx, S. W. Epp, Y. Zhong, A. Kuo, A. R. Balo, J. Soman, F. Schotte, H. T. Lemke, R. L. Owen, E. F. Pai, A. R. Pearson, J. S. Olson, P. A. Anfinrud, O. P. Ernst and R. J. Dwayne Miller, *Struct. Dyn.*, 2015, **2**, 054302–054317.
59. R. G. Sierra, H. Laksmono, J. Kern, R. Tran, J. Hattne, R. Alonso-Mori, B. Lassalle-Kaiser, C. Glockner, J. Hellmich, D. W. Schafer, N. Echols, R. J. Gildea, R. W. Grosse-Kunstleve, J. Sellberg, T. A. McQueen, A. R. Fry, M. M. Messerschmidt, A. Miahnahri, M. M. Seibert, C. Y. Hampton, D. Starodub, N. D. Loh, D. Sokaras, T. C. Weng, P. H. Zwart, P. Glatzel, D. Milathianaki, W. E. White, P. D. Adams, G. J. Williams, S. Boutet, A. Zouni, J. Messinger, N. K. Sauter, U. Bergmann, J. Yano, V. K. Yachandra and M. J. Bogan, *Acta Crystallogr., Sect. D*, 2012, **68**, 1584–1587.
60. D. Oberthuer, J. Knoska, M. O. Wiedorn, K. R. Beyerlein, D. A. Bushnell, E. G. Kovaleva, M. Heymann, L. Gumprecht, R. A. Kirian, A. Barty, V. Mariani, A. Tolstikova, L. Adriano, S. Awel, M. Barthelmess, K. Dorner, P. L. Xavier, O. M. Yefanov, D. R. James, G. Nelson, D. Wang,

G. Calvey, Y. Chen, A. Schmidt, M. Szczepek, S. Frielingsdorf, O. Lenz, E. Snell, P. J. Robinson, B. Šarler, G. Belšak, M. Maček, F. Wilde, A. Aquila, S. Boutet, M. Liang, M. S. Hunter, P. Scheerer, J. D. Lipscomb, U. Weierstall, R. D. Kornberg, J. C. H. Spence, L. Pollack, H. N. Chapman and S. Bajt, *Sci. Rep.*, 2017, **7**, 44628.

61. P. Roedig, H. M. Ginn, T. Pakendorf, G. Sutton, K. Harlos, T. S. Walter, J. Meyer, P. Fischer, R. Duman, I. Vartiainen, B. Reime, M. Warmer, A. S. Brewster, I. D. Young, T. Michels-Clark, N. K. Sauter, A. Kotecha, J. Kelly, D. J. Rowlands, M. Sikorsky, S. Nelson, D. S. Damiani, R. Alonso-Mori, J. Ren, E. E. Fry, C. David, D. I. Stuart, A. Wagner and A. Meents, *Nat. Methods*, 2017, **14**, 805–810.
62. B. Pedrini, C.-J. Tsai, G. Capitani, C. Padeste, M. S. Hunter, N. A. Zatsepin, A. Barty, W. H. Benner, S. Boutet, G. K. Feld, S. P. Hau-Riege, R. A. Kirian, C. Kupitz, M. M. Messerschmidt, J. I. Ogren, T. Pardini, B. Segelke, G. J. Williams, J. C. H. Spence, R. Abela, M. Coleman, J. E. Evans, G. Schertler, M. Frank and X.-D. Li, *Phil. Trans. R. Soc. B*, 2014, **369**, 20130500.
63. M. Frank, D. B. Carlson, M. S. Hunter, G. J. Williams, M. M. Messerschmidt, N. A. Zatsepin, A. Barty, W. H. Benner, K. Chu, A. T. Graf, S. P. Hau-Riege, R. A. Kirian, C. Padeste, T. Pardini, B. Pedrini, B. Segelke, M. M. Seibert, J. C. H. Spence, C.-J. Tsai, S. M. Lane, X.-D. Li, G. Schertler, S. Boutet, M. Coleman and J. E. Evans, *IUCrJ*, 2014, **1**, 305–317.
64. C. M. Casadei, C.-J. Tsai, A. Barty, M. S. Hunter, N. A. Zatsepin, C. Padeste, G. Capitani, W. H. Benner, S. Boutet, S. P. Hau-Riege, C. Kupitz, M. M. Messerschmidt, J. I. Ogren, T. Pardini, K. J. Rothschild, L. Sala, B. Segelke, G. J. Williams, J. E. Evans, X.-D. Li, M. Coleman, B. Pedrini and M. Frank, *IUCrJ*, 2018, **5**, 103–117.
65. U. Weierstall, *Phil. Trans. R. Soc. B*, 2014, **369**, 20130337.
66. C. E. Conrad, S. Basu, D. James, D. Wang, A. Schaffer, S. Roy-Chowdhury, N. A. Zatsepin, A. Aquila, J. Coe, C. Gati, M. S. Hunter, J. E. Koglin, C. Kupitz, G. Nelson, G. Subramanian, T. A. White, Y. Zhao, J. D. Zook, S. Boutet, V. Cherezov, J. C. H. Spence, R. Fromme, U. Weierstall and P. Fromme, *IUCrJ*, 2015, **2**, 421–430.
67. J. M. Martin-Garcia, C. E. Conrad, G. Nelson, N. Stander, N. A. Zatsepin, J. D. Zook, L. Zhu, J. Geiger, E. Chun, D. Kissick, M. C. Hilgart, C. Ogata, A. Ishchenko, N. Nagaratnam, S. Roy-Chowdhury, J. Coe, G. Subramanian, A. Schaffer, D. James, G. Ketwala, N. Venugopalan, S. Xu, S. Corcoran, D, Ferguson, U. Weierstall, J. C. H. Spence, V. Cherezov, P. Fromme, R. F. Fischetti and W. Liu, *IUCrJ*, 2017, **4**, 1–16.
68. M. Sugahara, E. Mizohata, E. Nango, M. Suzuki, T. Tanaka, T. Masuda, R. Tanaka, T. Shimamura, Y. Tanaka, C. Suno, K. Ihara, D. Pan, K. Kakinouchi, S. Sugiyama, M. Murata, T. Inoue, K. Tono, C. Song, J. Park, T. Kameshima, T. Hatsui, Y. Joti, M. Yabashi and S. Iwata, *Nat. Methods*, 2014, **12**, 61–63.

69. K. Pande, C. D. M. Hutchison, G. Groenhof, A. Aquila, J. S. Robinson, J. Tenboer, S. Basu, S. Boutet, D. P. DePonte, M. Liang, T. A. White, N. A. Zatsepin, O. M. Yefanov, D. Morozov, D. Oberthuer, C. Gati, G. Subramanian, D. James, Y. Zhao, J. Koralek, J. Brayshaw, C. Kupitz, C. Conrad, S. Roy-Chowdhury, J. D. Coe, M. Metz, P. L. Xavier, T. D. Grant, J. E. Koglin, G. Ketawala, R. Fromme, V. Srajer, R. Henning, J. C. H. Spence, A. Ourmazd, P. Schwander, U. Weierstall, M. Frank, P. Fromme, A. Barty, H. N. Chapman, K. Moffat, J. J. Van Thor and M. Schmidt, *Science*, 2016, **352**, 725–729.
70. P. Nogly, V. Panneels, G. Nelson, C. Gati, T. Kimura, C. Milne, D. Milathianaki, M. Kubo, W. Wu, C. Conrad, J. Coe, R. Bean, Y. Zhao, P. B. A. th, R. Dods, R. Harimoorthy, K. R. Beyerlein, J. Rheinberger, D. James, D. Deponte, C. Li, L. Sala, G. J. Williams, M. S. Hunter, J. E. Koglin, P. Berntsen, E. Nango, S. Iwata, H. N. Chapman, P. Fromme, M. Frank, R. Abela, S. Boutet, A. Barty, T. A. White, U. Weierstall, J. C. H. Spence, R. Neutze, G. Schertler and J. O. R. Standfuss, *Nat. Commun.*, 2016, **7**, 1–9.
71. J. R. Stagno, Y. Liu, Y. R. Bhandari, C. E. Conrad, S. Panja, M. Swain, L. Fan, G. Nelson, C. Li, D. R. Wendel, T. A. White, J. D. Coe, M. O. Wiedorn, J. Knoska, D. Oberthuer, R. A. Tuckey, P. Yu, M. Dyba, S. G. Tarasov, U. Weierstall, T. D. Grant, C. D. Schwieters, J. Zhang, A. R. Ferré-D'Amaré, P. Fromme, D. E. Draper, M. Liang, M. S. Hunter, S. Boutet, K. Tan, X. Zuo, X. Ji, A. Barty, N. A. Zatsepin, H. N. Chapman, J. C. H. Spence, S. A. Woodson and Y. X. Wang, *Nature*, 2017, **541**, 242–246.
72. C. Kupitz, J. L. Olmos Jr., M. Holl, L. Tremblay, K. Pande, S. Pandey, D. Oberthür, M. Hunter, M. Liang, A. Aquila, J. Tenboer, G. Calvey, A. Katz, Y. Chen, M. O. Wiedorn, J. Knoska, A. Meents, V. Majriani, T. Norwood, I. Poudyal, T. D. Grant, M. D. Miller, W. Xu, A. Tolstikova, A. Morgan, M. Metz, J. M. Martin-Garcia, J. D. Zook, S. Roy-Chowdhury, J. Coe, N. Nagaratnam, D. Meza, R. Fromme, S. Basu, M. Frank, T. A. White, A. Barty, S. Bajt, O. M. Yefanov, H. N. Chapman, N. A. Zatsepin, G. Nelson, U. Weierstall, J. C. H. Spence, P. Schwander, L. Pollack, P. Fromme, A. Ourmazd, G. N. Phillips Jr. and M. Schmidt, *Struct. Dyn.*, 2017, **4**, 044003.
73. J. Olmos, S. Pandey, J. M. Martin-Garcia, G. Calvey, A. Katz, J. Knoska, C. Kupitz, M. S. Hunter, M. Liang, D. Oberthuer, O. M. Yefanov, M. Wiedorn, M. Heymann, M. Holl, K. Pande, A. Barty, M. D. Miller, S. Stern, S. Roy-Chowdhury, J. D. Coe, N. Nagaratnam, J. D. Zook, T. Norwood, I. Poudyal, D. Xu, J. Koglin, M. Seaberg, Y. Zhao, S. Bajt, T. D. Grant, V. Mariani, G. Nelson, G. Subramanian, E. Bae, R. Fromme, R. Fung, P. Schwander, M. Frank, T. A. White, U. Weierstall, N. A. Zatsepin, J. C. H. Spence, P. Fromme, H. N. Chapman, L. Pollack, L. Tremblay, A. Ourmazd, G. N. Phillips Jr. and M. Schmidt, *bioRxiv*, 2017, 1–40.
74. D. Wang, U. Weierstall, L. Pollack and J. C. H. Spence, *J. Synchrotron Radiat.*, 2014, 1–3.

75. G. D. Calvey, A. M. Katz, C. B. Schaffer and L. Pollack, *Struct. Dyn.*, 2016, **3**, 054301.
76. K. Moffat, in *X-Ray Free Electron Lasers*, ed. U. Bergmann, V. Yachandra and J. Yano, Royal Society of Chemistry, Cambridge, 2017, pp. 105–127.
77. V. Srajer and M. Schmidt, *J. Phys. D: Appl. Phys.*, 2017, **50**, 373001.
78. K. Moffat, *Phil. Trans. R. Soc. B*, 2014, **369**, 20130568.
79. A. Aquila, M. S. Hunter, R. B. Doak, R. A. Kirian, P. Fromme, T. A. White, J. Andreasson, D. Arnlund, S. Bajt, T. R. M. Barends, M. Barthelmess, M. J. Bogan, C. Bostedt, H. Bottin, J. D. Bozek, C. Caleman, N. Coppola, J. Davidsson, D. P. DePonte, V. Elser, S. W. Epp, B. Erk, H. Fleckenstein, L. Foucar, M. Frank, R. Fromme, H. Graafsma, I. Grotjohann, L. Gumprecht, J. Hajdu, C. Y. Hampton, A. Hartmann, R. Hartmann, S. P. Hau-Riege, G. Hauser, H. Hirsemann, P. Holl, J. M. Holton, A. Homke, L. Johansson, N. Kimmel, S. Kassemeyer, F. Krasniqi, K.-U. Kühnel, M. Liang, L. Lomb, E. Malmerberg, S. Marchesini, A. V. Martin, F. R. N. C. Maia, M. M. Messerschmidt, K. Nass, C. Reich, R. Neutze, D. Rolles, B. Rudek, A. Rudenko, I. Schlichting, C. Schmidt, K. E. Schmidt, J. Schulz, M. M. Seibert, R. L. Shoeman, R. Sierra, H. Soltau, D. Starodub, F. Stellato, S. Stern, L. Struder, N. Timneanu, J. Ullrich, X. Wang, G. J. Williams, G. Weidenspointner, U. Weierstall, C. B. Wunderer, A. Barty, J. C. H. Spence and H. N. Chapman, *Opt. Express*, 2012, **20**, 2706–2716.
80. J. Tenboer, S. Basu, N. A. Zatsepin, K. Pande, D. Milathianaki, M. Frank, M. Hunter, S. Boutet, G. J. Williams, J. E. Koglin, D. Oberthuer, M. Heymann, C. Kupitz, C. Conrad, J. Coe, S. Roy-Chowdhury, U. Weierstall, D. James, D. Wang, T. D. Grant, A. Barty, O. M. Yefanov, J. Scales, C. Gati, C. Seuring, V. Srajer, R. Henning, P. Schwander, R. Fromme, A. Ourmazd, K. Moffat, J. J. Van Thor, J. C. H. Spence, P. Fromme, H. N. Chapman and M. Schmidt, *Science*, 2014, **346**, 1242–1246.
81. E. Nango, A. Royant, M. Kubo, T. Nakane, C. Wickstrand, T. Kimura, T. Tanaka, K. Tono, C. Song, R. Tanaka, T. Arima, A. Yamashita, J. Kobayashi, T. Hosaka, E. Mizohata, P. Nogly, M. Sugahara, D. Nam, T. Nomura, T. Shimamura, D. Im, T. Fujiwara, Y. Yamanaka, B. Jeon, T. Nishizawa, K. Oda, M. Fukuda, R. Andersson, P. Bath, R. Dods, J. Davidsson, S. Matsuoka, S. Kawatake, M. Murata, O. Nureki, S. Owada, T. Kameshima, T. Hatsui, Y. Joti, G. Schertler, M. Yabashi, A.-N. Bondar, J. Standfuss, R. Neutze and S. Iwata, *Science*, 2016, **354**, 1552–1557.
82. I. D. Young, M. Ibrahim, R. Chatterjee, S. Gul, F. D. Fuller, S. Koroidov, A. S. Brewster, R. Tran, R. Alonso-Mori, T. Kroll, T. Michels-Clark, H. Laksmono, R. G. Sierra, C. A. Stan, R. Hussein, M. Zhang, L. Douthit, M. Kubin, C. de Lichtenberg, L. V. Pham, H. Nilsson, M. H. Cheah, D. Shevela, C. Saracini, M. A. Bean, I. Seuffert, D. Sokaras, T.-C. Weng, E. Pastor, C. Weninger, T. Fransson, L. Lassalle, P. Bräuer, P. Aller, P. T. Docker, B. Andi, A. M. Orville, J. M. Glownia, S. Nelson, M. Sikorski,

D. Zhu, M. S. Hunter, T. J. Lane, A. Aquila, J. E Koglin, J. Robinson, M. Liang, S. Boutet, A. Y. Lyubimov, M. Uervirojnangkoorn, N. W. Moriarty, D. Liebschner, P. V. Afonine, D. G. Waterman, G. Evans, P. Wernet, H. Dobbek, W. I. Weis, A. T. Brunger, P. H. Zwart, P. D. Adams, A. Zouni, J. Messinger, U. Bergmann, N. K. Sauter, J. Kern, V. K. Yachandra and J. Yano, *Nature*, 2016, **540**, 453–457.

83. T. R. M. Barends, L. Foucar, A. Ardevol, K. Nass, A. Aquila, S. Botha, R. B. Doak, K. Falahati, E. Hartmann, M. Hilpert, M. Heinz, M. C. Hoffmann, J. Köfinger, J. E. Koglin, G. Kovacsova, M. Liang, D. Milathianaki, H. Lemke, J. Reinstein, C. M. Roome, R. L. Shoeman, G. J. Williams, I. Burghardt, G. Hummer, S. Boutet and I. Schlichting, *Science*, 2015, **350**, 445–450.
84. N. Coquelle, M. Sliwa, J. Woodhouse, G. Schirò, V. Adam, A. Aquila, T. R. M. Barends, S. Boutet, M. Byrdin, S. Carbajo, E. De la Mora, R. B. Doak, M. Feliks, F. Fieschi, L. Foucar, V. Guillon, M. Hilpert, M. S. Hunter, S. Jakobs, J. E. Koglin, G. Kovacsova, T. J. Lane, B. Lévy, M. Liang, K. Nass, J. Ridard, J. S. Robinson, C. M. Roome, C. Ruckebusch, M. Seaberg, M. Thepaut, M. Cammarata, I. Demachy, M. Field, R. L. Shoeman, D. Bourgeois, J.-P. Colletier, I. Schlichting and M. Weik, *Nat. Chem.*, 2017, **10**, 31–37.
85. M. Schmidt, *Adv. Condens. Matter Phys.*, 2013, **2013**, 1–10.
86. C. Kupitz, I. Grotjohann, C. E. Conrad, S. Roy-Chowdhury, R. Fromme and P. Fromme, *Phil. Trans. R. Soc. B*, 2014, **369**, 20130316.
87. D. Arnlund, L. C. Johansson, C. Wickstrand, A. Barty, G. J. Williams, E. Malmerberg, J. Davidsson, D. Milathianaki, D. P. DePonte, R. L. Shoeman, D. Wang, D. James, G. Katona, S. Westenhoff, T. A. White, A. Aquila, S. Bari, P. Berntsen, M. Bogan, T. B. van Driel, R. B. Doak, K. S. Kjær, M. Frank, R. Fromme, I. Grotjohann, R. Henning, M. S. Hunter, R. A. Kirian, I. Kosheleva, C. Kupitz, M. Liang, A. V. Martin, M. M. Nielsen, M. M. Messerschmidt, M. M. Seibert, J. Sjöhamn, F. Stellato, U. Weierstall, N. A. Zatsepin, J. C. H. Spence, P. Fromme, I. Schlichting, S. Boutet, G. Groenhof, H. N. Chapman and R. Neutze, *Nat. Methods*, 2014, **11**, 923–926.
88. M. Levantino, G. Schirò, H. T. Lemke, G. Cottone, J. M. Glownia, D. Zhu, M. Chollet, H. Ihee, A. Cupane and M. Cammarata, *Nat. Commun.*, 2015, 1–6.
89. L. Foucar, A. Barty, N. Coppola, R. Hartmann, P. Holl, U. Hoppe, S. Kassemeyer, N. Kimmel, J. Küpper, M. Scholz, S. Techert, T. A. White, L. Strüder and J. Ullrich, *Comput. Phys. Commun.*, 2012, **183**, 2207–2213.
90. D. Damiani, M. Dubrovin, I. Gaponenko, W. Kroeger, T. J. Lane, A. Mitra, C. P. O'Grady, A. Salnikov, A. Sanchez-Gonzalez, D. Schneider and C. H. Yoon, *J. Appl. Crystallogr.*, 2016, **49**, 672–679.
91. A. Barty, R. Kirian, F. R. N. C. Maia, M. Hantke, C. H. Yoon, T. A. White and H. Chapman, *J. Appl. Crystallogr.*, 2014, **47**, 1118–1141.
92. N. K. Sauter, J. Hattne, R. W. Grosse-Kunstleve and N. Echols, *Acta Crystallogr., Sect. D*, 2013, **69**, 1274–1282.

93. O. B. Zeldin, A. S. Brewster, J. Hattne, M. Uervirojnangkoorn, A. Y. Lyubimov, Q. Zhou, M. Zhao, W. I. Weis, N. K. Sauter and A. T. Brunger, *Acta Crystallogr., Sect. D*, 2015, **71**, 352–356.
94. B. J. Daurer, M. F. Hantke, C. Nettelblad and F. R. N. C. Maia, *J. Appl. Crystallogr.*, 2016, **49**, 1042–1047.
95. V. Mariani, A. Morgan, C. H. Yoon, T. J. Lane, T. A. White, C. O'Grady, M. Kuhn, S. Aplin, J. E. Koglin, A. Barty and H. N. Chapman, *J. Appl. Crystallogr.*, 2016, **49**, 1073–1080.
96. P. Hart, S. Boutet, G. Carini, A. Dragone, B. Duda, D. Freytag, G. Haller, R. Herbst, S. Herrmann, C. Kenney, J. Morse, M. Nordby, J. Pines, N. van Bakel, M. Weaver and G. J. Williams, *Nuclear Science Symposium & Medical Imaging Conference (NSS/MIC), IEEE*, 2012, 538–541.
97. T.-Y. Lan, J. L. Wierman, M. W. Tate, H. T. Philipp, V. Elser and S. M. Gruner, *J. Appl. Crystallogr.*, 2017, **50**, 985–993.
98. B. Henrich, J. Becker, R. Dinapoli, P. Goettlicher, H. Graafsma, H. Hirsemann, R. Klanner, H. Krueger, R. Mazzocco, A. Mozzanica, H. Perrey, G. Potdevin, B. Schmitt, X. Shi, A. K. Srivastava, U. Trunk and C. Youngman, *Nucl. Instrum. Methods Phys. Res., A*, 2010, **633**, 1–4.
99. T. Tschentscher, C. Bressler, J. Grünert, A. Madsen, A. Mancuso, M. Meyer, A. Scherz, H. Sinn and U. Zastrau, *Appl. Sci.*, 2017, **7**, 592.
100. T. A. White, R. A. Kirian, A. V. Martin, A. Aquila, K. Nass, A. Barty and H. N. Chapman, *J. Appl. Crystallogr.*, 2012, **45**, 335–341.
101. T. A. White, A. Barty, F. Stellato, J. M. Holton, R. A. Kirian, N. A. Zatsepin and H. N. Chapman, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2013, **69**, 1231–1240.
102. T. A. White, V. Mariani, W. Brehm, O. M. Yefanov, A. Barty, K. R. Beyerlein, F. Chervinskii, L. Galli, C. Gati, T. Nakane, A. Tolstikova, K. Yamashita, C. H. Yoon, K. Diederichs and H. N. Chapman, *J. Appl. Crystallogr.*, 2015, **49**, 680–689.
103. T. Nakane, Y. Joti, K. Tono, M. Yabashi, E. Nango, S. Iwata, R. Ishitani and O. Nureki, *J. Appl. Crystallogr.*, 2016, **49**, 1–7.
104. B. G. Abdallah, N. A. Zatsepin, S. Roy-Chowdhury, J. Coe, C. E. Conrad, K. Dorner, R. G. Sierra, H. P. Stevenson, F. Camacho-Alanis, T. D. Grant, G. Nelson, D. James, G. Calero, R. M. Wachter, J. C. H. Spence, U. Weierstall, P. Fromme and A. Ros, *Struct. Dyn*, 2015, **2**, 041719.
105. R. A. Kirian, X. Wang, U. Weierstall, K. E. Schmidt, J. C. H. Spence, M. Hunter, P. Fromme, T. A. White, H. N. Chapman and J. Holton, *Opt. Express*, 2010, **18**, 5713–5723.
106. C. Nave, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 1998, **54**, 848–853.
107. M. Uervirojnangkoorn, O. B. Zeldin, A. Y. Lyubimov, J. Hattne, A. S. Brewster, N. K. Sauter, A. T. Brunger and W. I. Weis, *eLife*, 2015, 1–29.
108. M. G. Rossmann, A. G. W. Leslie, S. S. Abdel-Meguid and T. Tsukihara, *J. Appl. Crystallogr.*, 1979, **12**, 570–581.
109. R. A. Kirian, T. A. White, J. M. Holton, H. N. Chapman, P. Fromme, A. Barty, L. Lomb, A. Aquila, F. R. N. C. Maia, A. V. Martin, R. Fromme,

X. Wang, M. S. Hunter, K. E. Schmidt and J. C. H. Spence, *Acta Crystallogr., Sect. A*, 2011, **67**, 131–140.
110. C. Li, K. Schmidt and J. C. H. Spence, *Struct. Dyn.*, 2015, **2**, 041714.
111. J. C. H. Spence, N. A. Zatsepin and C. Li, *Phil. Trans. R. Soc. B*, 2014, **369**, 20130325.
112. J. Hattne, N. Echols, R. Tran, J. Kern, R. J. Gildea, A. S. Brewster, R. Alonso-Mori, C. Glöckner, J. Hellmich, H. Laksmono, R. G. Sierra, B. Lassalle-Kaiser, A. Lampe, G. Han, S. Gul, D. DiFiore, D. Milathianaki, A. R. Fry, A. Miahnahri, W. E. White, D. W. Schafer, M. M. Seibert, J. E. Koglin, D. Sokaras, T.-C. Weng, J. Sellberg, M. J. Latimer, P. Glatzel, P. H. Zwart, R. W. Grosse-Kunstleve, M. J. Bogan, M. M. Messerschmidt, G. J. Williams, S. Boutet, J. Messinger, A. Zouni, J. Yano, U. Bergmann, V. K. Yachandra, P. D. Adams and N. K. Sauter, *Nat. Methods*, 2014, **11**, 545–548.
113. J. Hattne, N. Echols, R. Tran, J. Kern, R. J. Gildea, A. S. Brewster, R. Alonso-Mori, C. Glöckner, J. Hellmich, H. Laksmono, R. G. Sierra, B. Lassalle-Kaiser, A. Lampe, G. Han, S. Gul, D. DiFiore, D. Milathianaki, A. R. Fry, A. Miahnahri, W. E. White, D. W. Schafer, M. M. Seibert, J. E. Koglin, D. Sokaras, T.-C. Weng, J. Sellberg, M. J. Latimer, P. Glatzel, P. H. Zwart, R. W. Grosse-Kunstleve, M. J. Bogan, M. M. Messerschmidt, G. J. Williams, S. Boutet, J. Messinger, A. Zouni, J. Yano, U. Bergmann, V. K. Yachandra, P. D. Adams and N. K. Sauter, *Nat. Methods*, 2014, **11**, 545–548.
114. T. A. White, *Phil. Trans. R. Soc. B*, 2014, **369**, 20130330.
115. N. K. Sauter, J. Hattne, A. S. Brewster, N. Echols, P. H. Zwart and P. D. Adams, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2014, **70**, 3299–3309.
116. N. K. Sauter, *J. Synchrotron Radiat.*, 2015, **22**, 239–248.
117. H. M. Ginn, M. M. Messerschmidt, X. Ji, H. Zhang, D. Axford, R. J. Gildea, G. Winter, A. S. Brewster, J. Hattne, A. Wagner, J. M. Grimes, G. Evans, N. K. Sauter, G. Sutton and D. I. Stuart, *Nat. Commun.*, 2015, **6**, 1–8.
118. W. Kabsch, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2014, **70**, 2204–2216.
119. S. Botha, K. Nass, T. R. M. Barends, W. Kabsch, B. Latz, F. Dworkowski, L. Foucar, E. Panepucci, M. Wang, R. L. Shoeman, I. Schlichting and R. B. Doak, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2015, **71**, 387–397.
120. H. M. Ginn, A. S. Brewster, J. Hattne, G. Evans, A. Wagner, J. M. Grimes, N. K. Sauter, G. Sutton and D. I. Stuart, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2015, **71**, 1400–1410.
121. W. Brehm and K. Diederichs, *Crystallogr., Sect. D: Biol. Crystallogr.*, 2013, **70**, 101–109.
122. H. Liu and J. C. H. Spence, *IUCrJ*, 2014, 1–9.
123. G. M. Sheldrick, *Acta Crystallogr., Sect. A*, 2008, **64**, 112–122.
124. K. Yamashita, D. Pan, T. Okuda, M. Sugahara, A. Kodan, T. Yamaguchi, T. Murai, K. Gomi, N. Kajiyama, E. Mizohata, M. Suzuki, E. Nango, K. Tono, Y. Joti, T. Kameshima, J. Park, C. Song, T. Hatsui, M. Yabashi,

S. Iwata, H. Kato, H. Ago, M. Yamamoto and T. Nakatsu, *Sci. Rep.*, 2015, 1–10.

125. H. T. Philipp, M. Hromalik, M. W. Tate, L. J. Koerner and S. M. Gruner, *Nucl. Instrum. Methods Phys. Res. A*, 2011, **649**, 67–69.
126. K. Nass, A. Meinhart, T. R. M. Barends, L. Foucar, A. Gorel, A. Aquila, S. Botha, R. B. Doak, J. E. Koglin, M. Liang, R. L. Shoeman, G. Williams, S. Boutet and I. Schlichting, *IUCrJ*, 2016, **3**, 1–12.
127. K. Yamashita, N. Kuwabara, T. Nakane, T. Murai, E. Mizohata, M. Sugahara, D. Pan, T. Masuda, M. Suzuki, T. Sato, A. Kodan, T. Yamaguchi, E. Nango, T. Tanaka, K. Tono, Y. Joti, T. Kameshima, T. Hatsui, M. Yabashi, H. Manya, T. Endo, R. Kato, T. Senda, H. Kato, S. Iwata, H. Ago, M. Yamamoto, F. Yumoto and T. Nakatsu, *IUCrJ*, 2017, **4**, 639–647.
128. M. S. Hunter, C. H. Yoon, H. DeMirci, R. G. Sierra, E. H. Dao, R. Ahmadi, F. Aksit, A. L. Aquila, H. Ciftci, S. Guillet, M. J. Hayes, T. J. Lane, M. Liang, U. Lundström, J. E. Koglin, P. Mgbam, Y. Rao, L. Zhang, S. Wakatsuki, J. M. Holton and S. Boutet, *Nat. Commun.*, 2016, **7**, 1–5.
129. T. Nakane, C. Song, M. Suzuki, E. Nango, J. Kobayashi, T. Masuda, S. Inoue, E. Mizohata, T. Nakatsu, T. Tanaka, R. Tanaka, T. Shimamura, K. Tono, Y. Joti, T. Kameshima, T. Hatsui, M. Yabashi, O. Nureki, S. Iwata and M. Sugahara, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2015, **71**, 2519–2525.
130. T. Nakane, S. Hanashima, M. Suzuki, H. Saiki, T. Hayashi, K. Kakinouchi, S. Sugiyama, S. Kawatake, S. Matsuoka, N. Matsumori, E. Nango, J. Kobayashi, T. Shimamura, K. Kimura, C. Mori, N. Kunishima, M. Sugahara, Y. Takakyu, S. Inoue, T. Masuda, T. Hosaka, K. Tono, Y. Joti, T. Kameshima, T. Hatsui, M. Yabashi, T. Inoue, O. Nureki, S. Iwata, M. Murata and E. Mizohata, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, 13039–13044.
131. A. Batyuk, L. Galli, A. Ishchenko, G. W. Han, C. Gati, P. A. Popov, M.-Y. Y. Lee, B. Stauch, T. A. White, A. Barty, A. Aquila, M. S. Hunter, M. Liang, S. Boutet, M. Pu, Z.-J. Liu, G. Nelson, D. James, C. Li, Y. Zhao, J. C. H. Spence, W. Liu, P. Fromme, V. Katritch, U. Weierstall, R. C. Stevens and V. Cherezov, *Sci. Adv.*, 2016, **2**, e1600292.
132. Y. Fukuda, K. M. Tse, T. Nakane, T. Nakatsu, M. Suzuki, M. Sugahara, S. Inoue, T. Masuda, F. Yumoto, N. Matsugaki, E. Nango, K. Tono, Y. Joti, T. Kameshima, C. Song, T. Hatsui, M. Yabashi, O. Nureki, M. E. P. Murphy, T. Inoue, S. Iwata and E. Mizohata, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, 2928–2933.
133. J.-P. Colletier, M. R. Sawaya, M. Gingery, J. A. Rodriguez, D. Cascio, A. S. Brewster, T. Michels-Clark, R. H. Hice, N. Coquelle, S. Boutet, G. J. Williams, M. M. Messerschmidt, D. P. DePonte, R. G. Sierra, H. Laksmono, J. E. Koglin, M. S. Hunter, H.-W. Park, M. Uervirojnangkoorn, D. K. Bideshi, A. T. Brunger, B. A. Federici, N. K. Sauter and D. S. Eisenberg, *Nature*, 2016, **539**, 43–47.

134. A. Marinelli, D. Ratner, A. A. Lutman, J. Turner, J. Welch, F. J. Decker, H. Loos, C. Behrens, S. Gilevich, A. A. Miahnahri, S. Vetter, T. J. Maxwell, Y. Ding, R. Coffee, S. Wakatsuki and Z. Huang, *Nat. Commun.*, 2015, **6**, 1–6.
135. T. Hara, Y. Inubushi, T. Katayama, T. Sato, H. Tanaka, T. Tanaka, T. Togashi, K. Togawa, K. Tono, M. Yabashi and T. Ishikawa, *Nat. Commun.*, 2013, **4**, 1–5.
136. A. Gorel, K. Motomura, H. Fukuzawa, R. B. Doak, M. L. Grünbein, M. Hilpert, I. Inoue, M. Kloos, G. Kovacsova, E. Nango, K. Nass, C. M. Roome, R. L. Shoeman, R. Tanaka, K. Tono, Y. Joti, M. Yabashi, S. Iwata, L. Foucar, K. Ueda, T. R. M. Barends and I. Schlichting, *Nat. Commun.*, 2017, 1–7.
137. F. R. N. C. Maia, *Nat. Methods*, 2012, **9**, 854–855.
138. T. A. White, A. Barty, W. Liu, A. Ishchenko, H. Zhang, C. Gati, N. A. Zatsepin, S. Basu, D. Oberthür, M. Metz, K. R. Beyerlein, C. H. Yoon, O. M. Yefanov, D. James, D. Wang, M. M. Messerschmidt, J. E. Koglin, S. Boutet, U. Weierstall and V. Cherezov, *Sci. Data*, 2016, **3**, 160057.
139. C. H. Yoon, H. DeMirci, R. G. Sierra, E. H. Dao, R. Ahmadi, F. Aksit, A. L. Aquila, A. Batyuk, H. Ciftci, S. Guillet, M. J. Hayes, B. Hayes, T. J. Lane, M. Liang, U. Lundström, J. E. Koglin, P. Mgbam, Y. Rao, T. Rendahl, E. Rodriguez, L. Zhang, S. Wakatsuki, S. Boutet, J. M. Holton and M. S. Hunter, *Sci. Data*, 2017, **4**, 170055.
140. A. Munke, J. Andreasson, A. Aquila, S. Awel, K. Ayyer, A. Barty, R. J. Bean, P. Berntsen, J. Bielecki, S. Boutet, M. Bucher, H. N. Chapman, B. J. Daurer, H. DeMirci, V. Elser, P. Fromme, J. Hajdu, M. F. Hantke, A. Higashiura, B. G. Hogue, A. Hosseinizadeh, Y. Kim, R. A. Kirian, H. K. N. Reddy, T.-Y. Lan, D. S. D. Larsson, H. Liu, N. D. Loh, F. R. N. C. Maia, A. P. Mancuso, K. Mühlig, A. Nakagawa, D. Nam, G. Nelson, C. Nettelblad, K. Okamoto, A. Ourmazd, M. Rose, G. van der Schot, P. Schwander, M. M. Seibert, J. A. Sellberg, R. G. Sierra, C. Song, M. Svenda, N. Timneanu, I. A. Vartanyants, D. Westphal, M. O. Wiedorn, G. J. Williams, P. L. Xavier, C. H. Yoon and J. D. Zook, *Sci. Data*, 2016, **3**, 160064.
141. X. E. Zhou, X. Gao, A. Barty, Y. Kang, Y. He, W. Liu, A. Ishchenko, T. A. White, O. M. Yefanov, G. W. Han, Q. Xu, P. W. de Waal, K. M. Suino-Powell, S. Boutet, G. J. Williams, M. Wang, D. Li, M. Caffrey, H. N. Chapman, J. C. H. Spence, P. Fromme, U. Weierstall, R. C. Stevens, V. Cherezov, K. Melcher and H. E. Xu, *Sci. Data*, 2016, **3**, 160021.
142. J. L. Wierman, T. Y. Lan, M. W. Tate, H. T. Philipp, V. Elser and S. M. Gruner, *IUCrJ*, 2016, 1–8.
143. J. C. H. Spence, R. A. Kirian, X. Wang, U. Weierstall, K. E. Schmidt, T. A. White, A. Barty, H. N. Chapman, S. Marchesini and J. Holton, *Opt. Express*, 2011, **19**, 1–8.
144. R. A. Kirian, R. J. Bean, K. R. Beyerlein, M. Barthelmess, C. H. Yoon, F. Wang, F. Capotondi, E. Pedersoli, A. Barty and H. N. Chapman, *Phys. Rev. X*, 2015, **5**, 011015.

145. J. Chen, J. C. H. Spence and R. P. Millane, *Acta Crystallogr., Sect. A*, 2014, **70**, 154–161.
146. R. A. Kirian, R. J. Bean, K. R. Beyerlein, O. M. Yefanov, T. A. White, A. Barty and H. N. Chapman, *Phil. Trans. R. Soc. B*, 2014, **369**, 20130331.
147. H. N. Chapman and P. Fromme, *Curr. Opin. Struct. Biol.*, 2017, **45**, 170–177.
148. K. Ayyer, H. T. Philipp, M. W. Tate, J. L. Wierman, V. Elser and S. M. Gruner, *IUCrJ*, 2015, **2**, 29–34.
149. Henry N Chapman, O. M. Yefanov, K. Ayyer, T. A. White, A. Barty, A. Morgan, V. Mariani, D. Oberthuer and K. Pande, *arXiv*, 2017, 1–34.
150. J. C. H. Spence, U. Weierstall, T. T. Fricke, R. M. Glaeser and K. H. Downing, *J. Struct. Biol.*, 2003, **144**, 209–218.
151. H. He and W. P. Su, *Acta Crystallogr., Sect. A*, 2015, **71**, 92–98.
152. A. Classen, K. Ayyer, H. N. Chapman, R. Röhlsberger and J. von Zanthier, *Phys. Rev. Lett.*, 2017, **119**, 1839.
153. P. Nogly, D. James, D. Wang, T. A. White, N. A. Zatsepin, A. Shilova, G. Nelson, H. Liu, L. Johansson, M. Heymann, K. Jaeger, M. Metz, C. Wickstrand, W. Wu, P. Bath, P. Berntsen, D. Oberthuer, V. Panneels, V. Cherezov, H. N. Chapman, G. Schertler, R. Neutze, J. C. H. Spence, I. Moraes, M. Burghammer, J. Standfuss and U. Weierstall, *IUCrJ*, 2015, **2**, 168–176.
154. F. Stellato, D. Oberthur, M. Liang, R. Bean, C. Gati, O. M. Yefanov, A. Barty, A. Burkhardt, P. Fischer, L. Galli, R. A. Kirian, J. Meyer, S. Panneerselvam, C. H. Yoon, F. Chervinskii, E. Speller, T. A. White, C. Betzel, A. Meents and H. N. Chapman, *IUCrJ*, 2014, 1–9.
155. S. Botha, K. Nass, T. R. M. Barends, W. Kabsch, B. Latz, F. Dworkowski, L. Foucar, E. Panepucci, M. Wang, R. L. Shoeman, I. Schlichting and R. B. Doak, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2015, **71**, 387–397.
156. J. Standfuss and J. C. H. Spence, *IUCrJ*, 2017, **4**, 100–101.

Subject Index

acoustic droplet ejection (ADE), 9
acoustic waves, 19
American National Standards Institute (ANSI), 7
angiotensin II (AngII), 44
anisotropic data and crystal twinning
  experimental
    protein crystallization and derivatization, 142
    protein expression and purification, 141
    X-ray diffraction data collection and analysis, 142
  results and discussion
    attempts in experimental phasing, 151–152
    data collection and radiation damage, 142–143
    data reduction and anisotropy, 143–144
    difference Fourier analysis, 149–151
    ligand-binding site, 153–154
    molecular replacement and structure solution, 147–149
    NorM-NG structures, 152–153
    twinning detection and analysis, 144–147
ANSI. *See* American National Standards Institute (ANSI)

COC. *See* cyclic olefin copolymer (COC)
concentric hydrophilic–hydrophobic patterned rings, 11
COP. *See* cyclic olefin polymer (COP)
COP9 signalosome (CSN), 157–158
cryo-cooled single crystals, 55, 56
cryo-temperature data collection, 91
crystal characterisation, 161–165
crystallization, 159–161
CSN4, 165
cyclic olefin copolymer (COC), 10
cyclic olefin polymer (COP), 17

data collection and processing, 159–161
data quality analysis
  accuracy *versus* precision, 118–119
  estimating errors
    data completeness, 130
    estimation of $\sigma(I_{hkl})$, 125–126
    ISa (asymptotic signal-to-noise ratio), 126
    R_{d}, R_{cum} and *B factor* (radiation damage), 129–130
    R_{merge}, R_{sym} and R_{meas} (unmerged data), 126–127
    $R_{mrgd\text{-}I}$, $R_{p.i.m.}$, R_{anom} and $CC_{1/2}$ (merged data), 127–129

data quality analysis (*continued*)
merged *versus* unmerged data, 118–119
sources of error, 119–120
outliers, 123–124
radiation damage, 124–125
random error, 120–121
systematic error, 121–123
use of metrics
BLEND, 130–131
collecting and analysing data, 134–137
high-resolution cut-off, 132–134
linking data and model quality, 131–132
merging multiple crystals and/or data sets, 130–131
rogue data sets, 131–132
DEER. *See* double electron–electron resonance (DEER) spectrometry
difference Fourier analysis, 149–151
double electron–electron resonance (DEER) spectrometry, 45
double-sandwich methods, 15

electron microscopy (EM), 45
ergotamine (ERG), 42
European Molecular Biology Laboratory (EMBL), 10
European Synchrotron Radiation Facility (ESRF), 5

free interface diffusion (FID), 14

G protein-coupled receptors (GPCRs)
milestones
large-scale production, 31–37
receptor constructs, crystallization of, 38–39
process overview, 30–31
SFX data collection
data processing, 41–42
experimental layout, 39
LCP injector and crystal delivery, 39–41
XFELs, 42–46

heavy-atom-soaked derivative crystals, 165–166
Henderson limit, 4
hierarchical cluster analysis (HCA), 57
hydrogen/deuterium exchange (HDX) mass spectrometery, 45

in situ plate screening, 7
in situ X-ray crystallography
challenges of, 4–5
enabling technologies, 5–7
goals of, 3–4
liquid manipulation methods, 19
scale reduction and microfluidics
microfluidic methods for, 13–15
small formats, 11–13
synchrotron, standard SBS plates
automation and pipeline integration, 10–11
development history, 7–9
facilities, 6
plate handling hardware, 9
plate optimization for, 9–10
thin-film sandwiches, 15–19
I-SAD/I-SIRAS solutions, Mesh&Collect for, 68–69
isomorphous replacement (IR) technique, 55

lipid cubic phase (LCP), 2

macromolecular assemblies
crystal characterisation, 161–165
crystallization, 159–161
CSN4, 165
CSN conformational dynamics, 174–176
data collection and processing, 159–161
heavy-atom-soaked derivative crystals, 165–166
initial model building, 170–171
initial phasing, 166–169
model completion, 171–174
selenomethionine phasing, 169–170
subunit identification, 169–170
macromolecular crystallography (MX)
ccCluster, 75
DOZOR algorithm, 59
genetic algorithms, 75–77
fitness evaluation, 77–79
GA optimisation, 79
grouping partial data, 77
LUX, 79–80
hierarchical cluster analysis (HCA), 59–60
pitfalls of, 69–75
MeshBest, 80–82
NarQ crystals, 82
thaumatin crystals, mishmash of, 82–83
Mesh&Collect experiment, 60–61
crystal structures, 61–63
de novo structure solution, 64–68
multi-crystal data collection, 63–64
mass spectrometery, 45
Mesh&Collect pipeline. *See also* macromolecular crystallography (MX)
de novo structure solution, 64–68
I-SAD/I-SIRAS solutions, 68–69
in practice, 60–61
at room temperature, 68–69
thaumatin, crystal structure of, 64–68
MiTeGen, 11

NorM-NG structures, 152–153

δ-opioid receptor (DOR), 44
μ-opioid receptor (MOR), 44

phase problem, 88
primary photoelectron, 90
protein crystallization and derivatization, 142
protein expression and purification, 141

quantum mechanical tunneling, 91

RADDOSE, 98, 99
radiation damage
absorbed dose and dose limits, 96–102
chromophore containing Proteins, 102
global and specific radiation damage effects
at 100 K and below, 92–96
above 100 K and at room temperature, 102–103
incident beam energy at 100 K, 105–106
induced phasing, 104
macromolecular function, 103–104
minimization, 106–108
predominant classes of, 92
serial femtosecond crystallography at XFELs, 108–109
X-ray photons, 90–92
Roadrunner goniometer, 190

serial femtosecond crystallography (SFX), 3, 55
data analysis, 194–206

serial femtosecond crystallography (SFX) (*continued*)
data collection
data processing, 41–42
experimental layout, 39
LCP injector and crystal delivery, 39–41
membrane proteins, 30
potential power of, 39–41
early achievements, 187–188
experimental setup, 185–187
G protein-coupled receptors (GPCRs)
data processing, 41–42
experimental layout, 39
LCP injector and crystal delivery, 39–41
principle for, 29
sample delivery and data collection rates, 189–192
serial millisecond crystallography (SMX), 186
single wavelength anomalous dispersion (SAD), 99
Society for Biomolecular Screening (SBS), 7
Society for Laboratory Automation and Screening (SLAS), 7
sulfur single wavelength anomalous dispersion (S-SAD), 66, 204
Swiss Light Source (SLS), 10–11
X-ray free electron lasers (XFELs), 2, 55
continuous diffuse scattering, 206–207
incoherent diffractive imaging, 208
nanocrystal shape transform phasing, 206
radiation damage, 184–185
serial femtosecond crystallography
data analysis, 194–206
early achievements, 187–188
experimental setup, 185–187
sample delivery and data collection rates, 189–192
single-layer 2D crystals, 207–208
sparse crystal pattern indexing, 206
time-resolved serial femtosecond crystallography, 192–193
mix-and-inject serial femtosecond crystallography, 194
pump probe serial femtosecond crystallography, 193–194